国際会議場/3F
Day2 2025年7月31日(木) 15:45～17:00(1時間15分)

JANOG56

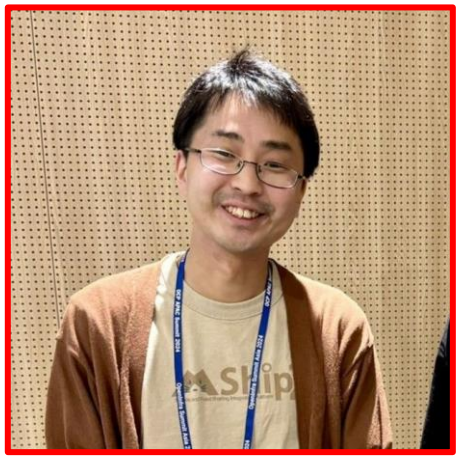
プライベートクラウドのデータセンターネットワークを作り直して得た知見の共有

KDDI株式会社

2025-07-31

Tomorrow, Together
KDDI

自己紹介



KDDI ネットワーク開発本部
キャリアクラウド開発部

辻　広志
Hiroshi Tsuji

出身：兵庫県三田市
職業：
手のひらネットワーク機器互換機
作成芸人

KDDI ネットワーク開発本部
キャリアクラウド開発部

深牧　紀彦
Toshihiko Fukamaki

出身： 島根県 松江市
担当： OpenStack環境の構築/運用
地元松江でのJANOGに参加するこ
とが出来て大変光栄です

KDDI ネットワーク開発本部
キャリアクラウド開発部

寺澤　大智
Daichi Terazawa

出身：兵庫県川西市
JANOG参加という貴重な機会を
頂きありがとうございます！

プライベートクラウドの紹介

タイトルにある「プライベートクラウド」のご紹介

Snapshot of 27-June 2025

MShip3 is OpenStack based KDDI's Private Cloud

Target Tenant: All of KDDI's network system (Mobile/Fixed Core, ISP, OSS/BSS)

Scale(Snapshot) :
- 3,500+ Hypervisors
- 350+ Storage servers
- 15,000+ VMs / 230,000+ vCPUs running
- 160+ Shift on Stack Clusters(4,000+ VMs/100,000+ vCPUs)
- 4PB Storage Used

Capacity(Snapshot): 563,000+ vCPU / Storage 35+ PB

タイトルにある「データセンターネットワーク」のご紹介

MShip3 using KDDI's DC Hardware platform named "KYANOS"

MShip3
Mobile and Fixed Sharing Integration Platform

Internal use,
Private Cloud

4,500+ servers

BareMetal K8s

特定ベンダ専用

1,000+ servers

VMware

エンタープライズ用途

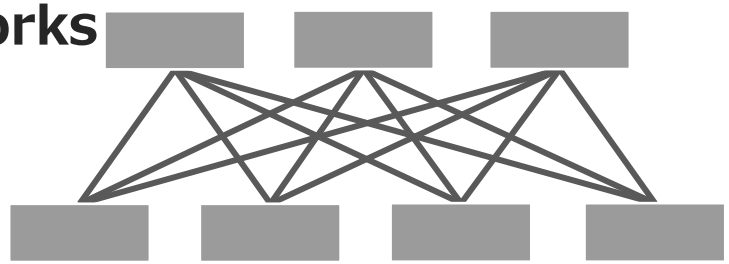
VDI

Virtual desktop
infra

KYANOS = KDDI's HaaS

Physical Servers
6,000+ servers

Datacenter
Networks



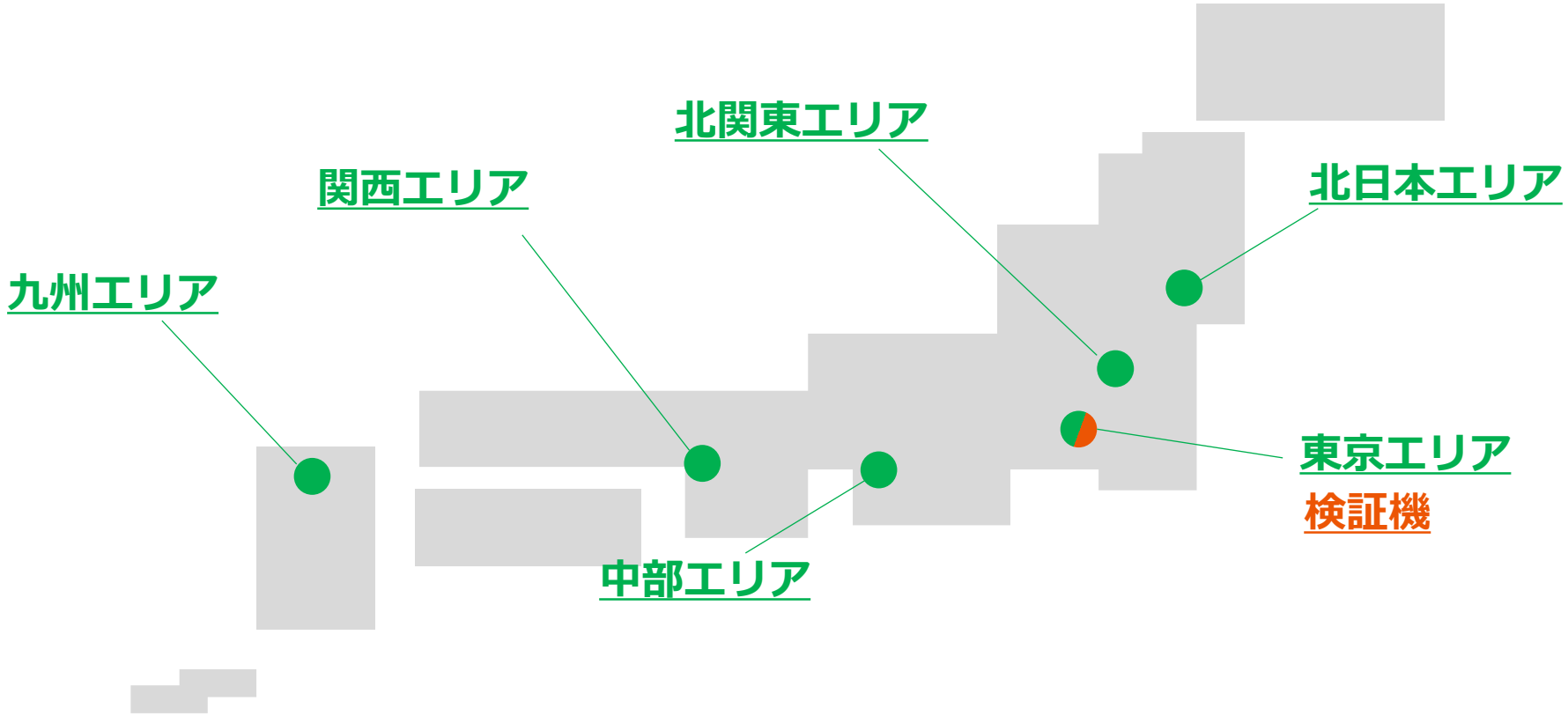
Tomorrow. Together
KDDI

Overview of MShip3 deployments

Snapshot of May 2025

6 Locations / 2 Global Clusters / 9 Regional Clusters

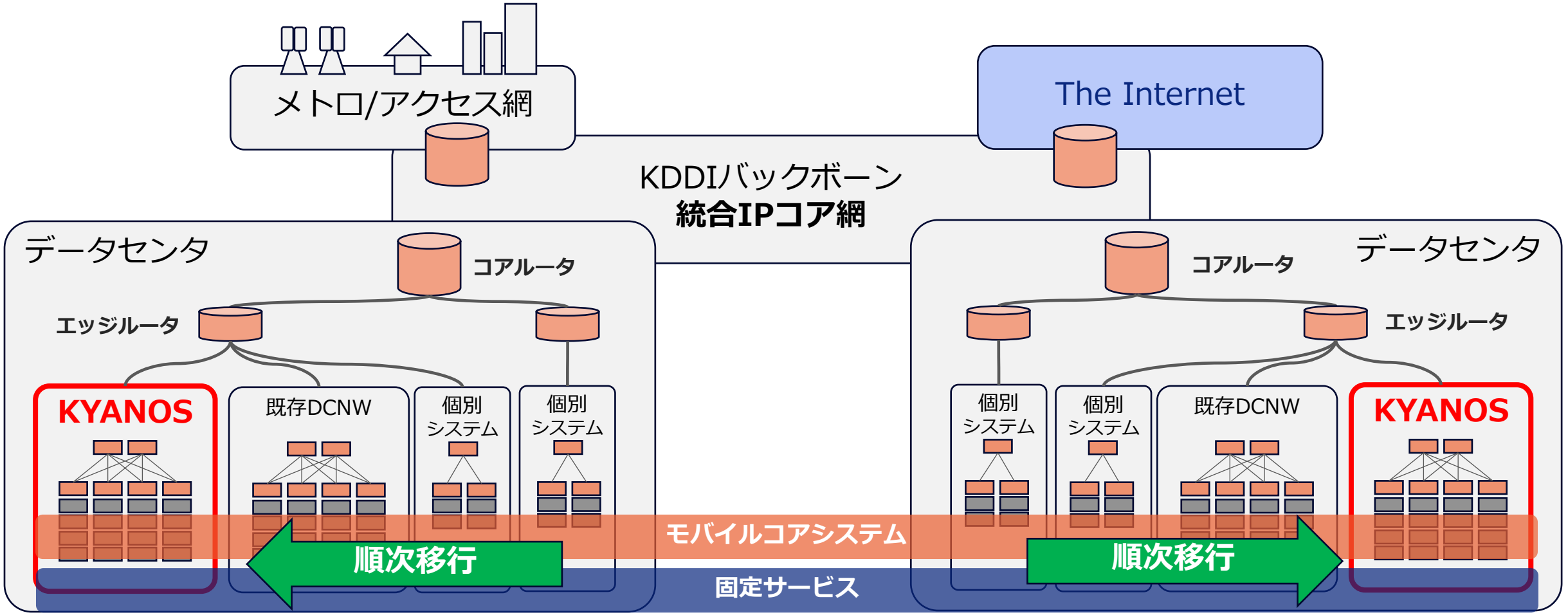
Each region has One OpenStack/Ceph Control-planes



Total 3,500+ Hypervisors

KDDI網内のおけるKYANOS/MShip3の立ち位置

KDDIのバックボーンのエッジルータ配下には様々なサービスを提供するシステムを収容
KYANOSはこれらの既存システムを仮想化して収容していくためにサーバとNWを提供



新しいデータセンターネットワークの設計コンセプト

これまでのKDDIのデータセンタインフラの課題

機能提供ベンダの仮想化基盤を採用

- アプリ毎に仮想化レイヤの設計が異なる
- 基盤側は仮想化レイヤ・アプリケーション毎に個別収容対応を実施

既存データセンタネットワークの構成

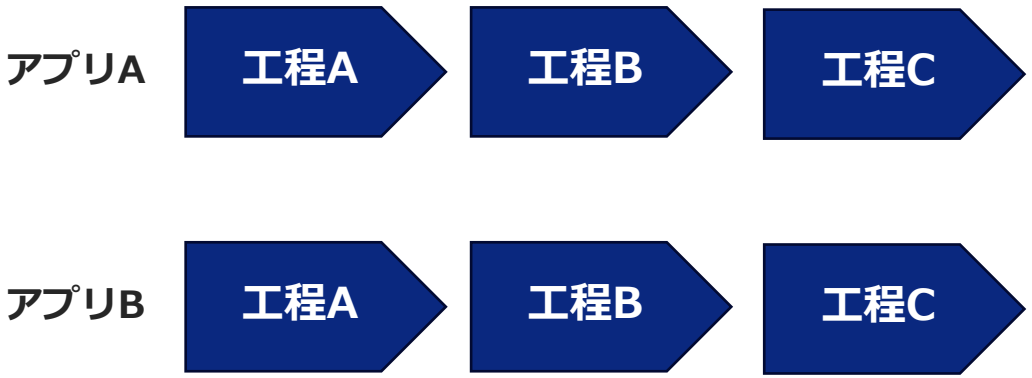
基盤チーム サービス提供箇所

過去の課題（設計面）

仮想化レイヤ・アプリケーション毎に個別対応を実施

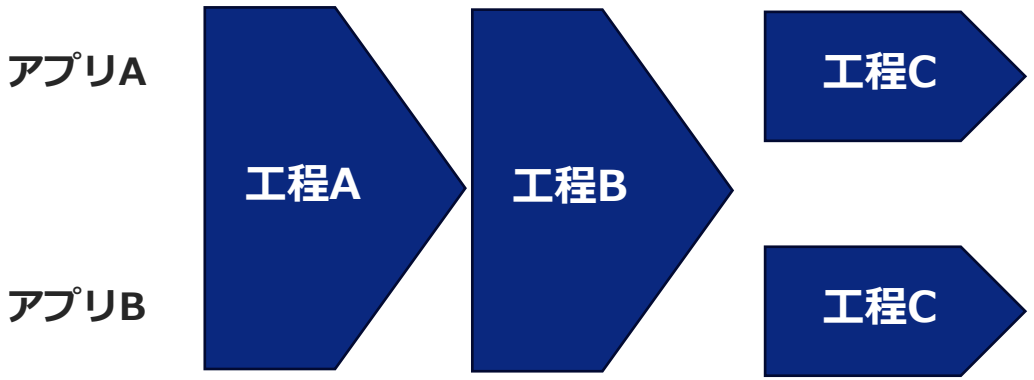
・新規アプリ収容毎に設計対応が必要（場合によってはHW調達も必要）
⇒基盤担当の稼働増加
⇒提供までのリードタイム増加

こうなっている



NW要件定義～導入までの工程をアプリ毎に実施

こうだったら嬉しい



共通化可能な工程はまとめて実施、共通化不可工程のみアプリ毎に実施

過去の課題（運用面）

物理・論理設計が異なることによる運用負荷の増大

	物理収容方式	L2	L3	L4-L7	利用VRF ・・・
アプリA	ACT-SBY	LLDP	BGP/Static	・・・	VRF1,3利用
アプリB	Port Chanel	-	OSPF	PBR	VRFA,C利用

・
・
・

トラシュー対応時

- ✕ アプリ毎に設計が異なり設計パターンが膨大
  ⇒逐一設計書の確認からの対応が必要
- ✕ ドキュメント整理が追い付いていないパターンも…
  ⇒ドキュメント確認と合わせて実機上での確認も必要

作業対応時(Version Up等)

- ✕ 検証パターンが膨大な事によるSI費用が増大する
  （〇億/年くらい必要なことも、、、）
- ✕ 設計パターン数が多すぎて、自動化できない or
  自動化してもリターンがコストに見合わない

下がっていく抽象度。上がる複雑化。

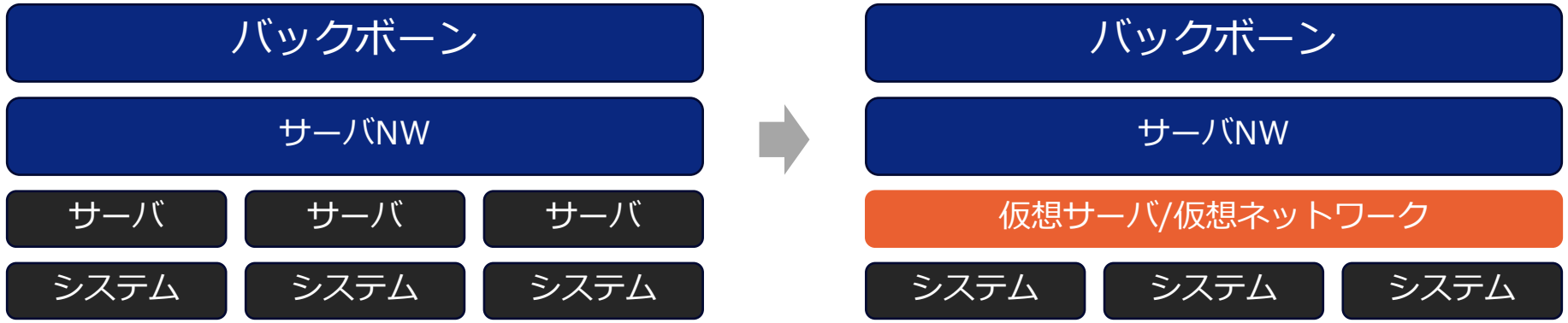
コスト削減、投資・体制の効率化のためにネットワーク設備の効率化を推進

網(KDDI)と設備(ベンダ)とのインタフェースはレイヤが下がるごとに複雑度が上昇。

「共通仮想化基盤」構想の登場@2021

ネットワークを含めて様々なインフラ上の課題を解決するために
「共通仮想化基盤」の検討が進めれれた

- **林立した仮想化システムを統一化する**
- **効率的に運用するために一つの仮想化基盤に集約する**
- **高いネットワーク側の負荷を仮想化やソフトウェア化での緩和を目指す**
  ➡共通化で下がってしまった抽象度を仮想化で再度上げる

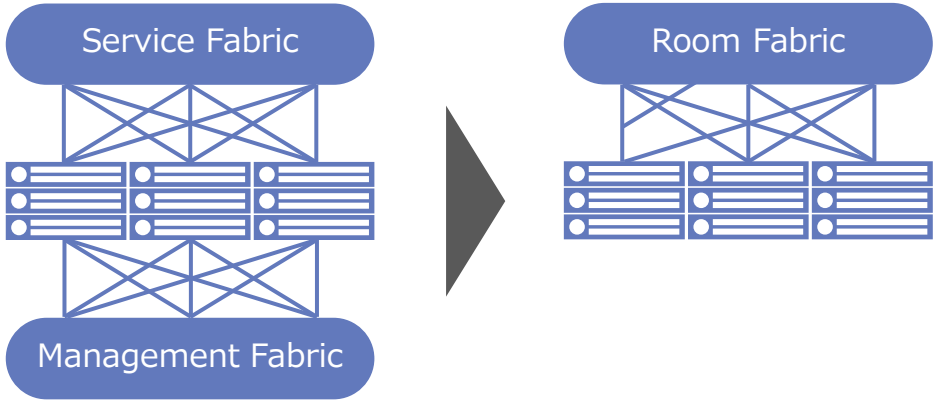


「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化①

ポリシーを現行から見直しシンプルさと高可用性を両立する

	設置単位	DCNWの大きさ
As is	用途+エリア	大きい
To be	部屋	小さい

DCNW設置ポリシーを用途別/エリア単位からを部屋単位へ



DCNWの一面化

得られる効果

- コスト削減（長距離伝送用トランシーバやTIE/線路が不要）
- 枯渇するDCNW側のリソース(設定上限値等)の削減
- 部屋間での共有リソース削減による可用性の向上
  （機械室内火災・発煙やSpineスイッチの間欠故障等）

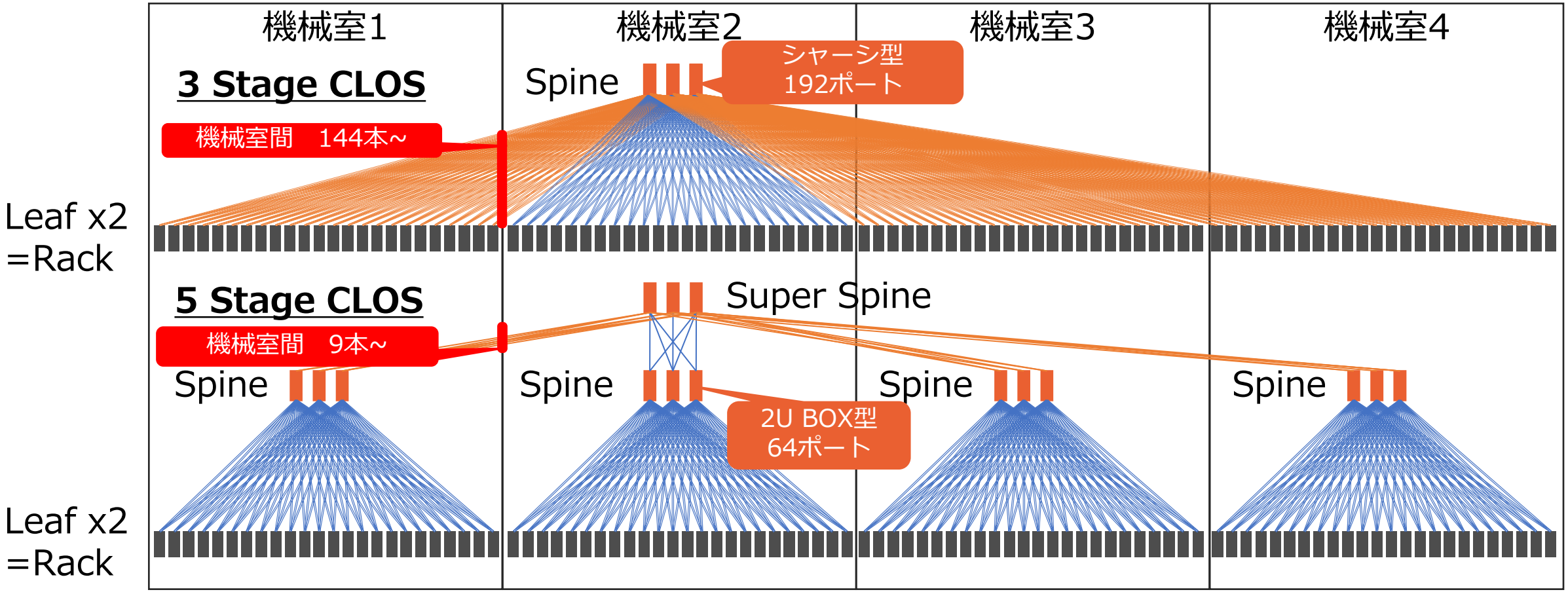
得られる効果

- コスト削減（部屋単位の設置で増えた分を削減）
- 故障時の動作のシンプル化
  （IaaSの管理系が切れた場合サービス系も切れる）
- 自動化のシンプル化
  （管理対象を一つのAZに対して一つにできる）

「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化②

5stage化によるコスト抑制

- スケーラビリティが向上し、Spineへの初期投資抑制、機械室間のファイバ削減が可能

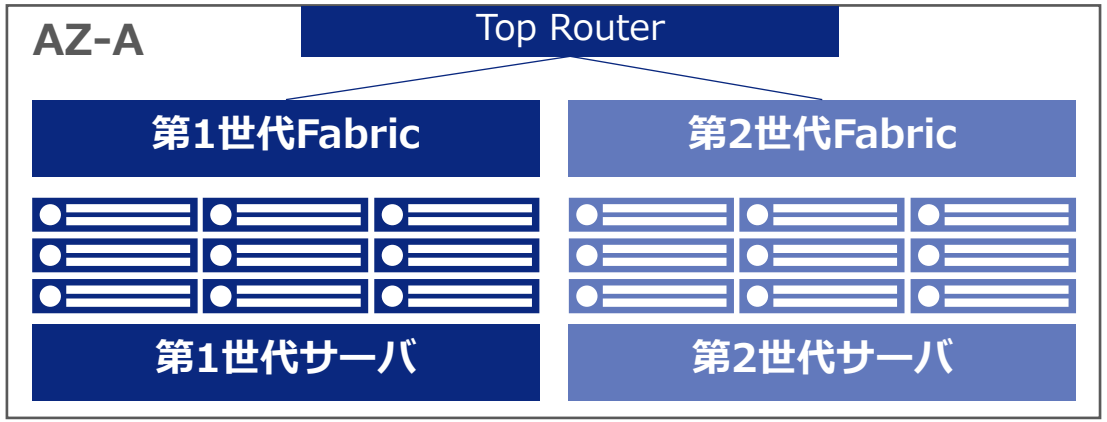


「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化③

データセンタネットワーク製品の流動性

既存ソリューションの課題

- 論理リソース最大数(設定項目等)の制限事項が多く、SI費やライセンス費など総合的に高額
- ホワイトボックススイッチ等を含めた第三のソリューションを検討したいが、KDDI側の検討リソース不足



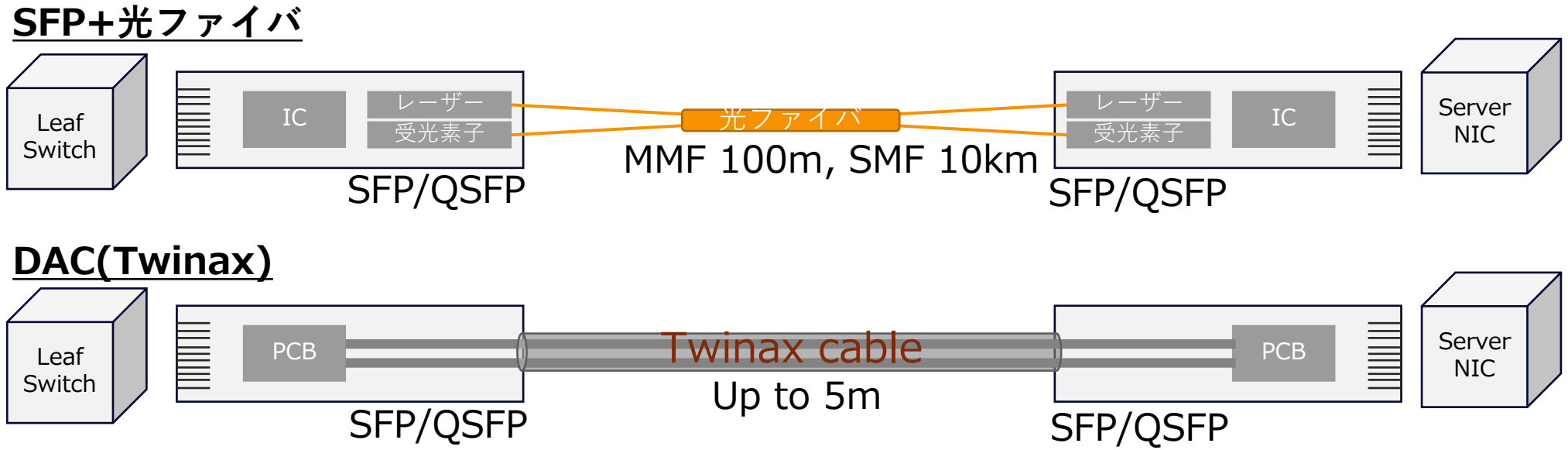
- **初期は既存製品のナレッジを活かして検討をスタートさせ、第二世代サーバのタイミングでFabric側の最適化検討を実施。**
- サーバ世代ごとにDCNWも入替えることでDCNWに流動性を持たせる

「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化④-1

DACケーブルの採用

サーバ~Leaf間は従来の光ファイバからDACに変更

- O/E変換を排し故障要素削減(別案件で導入しこれまで運用中の故障無し)
- NW接続コスト削減(O/E変換が不要な為、安価)
- サーバ~Leaf間の接続最適化(ラック内配線であれば十分な距離)



「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化④-2

Breakout Cableの採用とQSFP-56 NICの採用

Leafスイッチを400G化し、Breakoutにすることで高密度収容を実現

- 採用したNICはNFベンダ、基盤ベンダ共にドライバ対応済み
- スイッチメーカ純正品に4x100G-400G Breakout DACが無かったため、NICベンダ製ケーブルを採用
- 当該NICベンダのDACとスイッチメーカのスイッチ（今回採用とは異機種）の接続は別案件で実施済

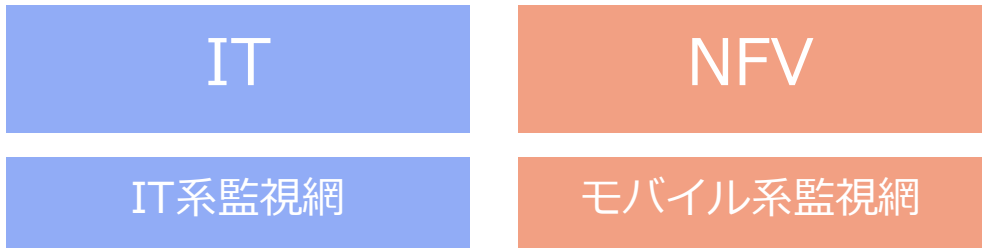
Source	Target	DACでの使用
1x 400G-QSFP56-DD(50G PAM4)	1x 400G-QSFP56-DD(50G PAM4)	**Passive**
	2x 200G-QSFP56 (50G PAM4)	**Passive**
	4x 100G-QSFP56 (50G PAM4)	**Passive**
	4x 100G-QSFP28(25G PAM2-NRZ)	**Active**

☞ **QSFP-56 NICの採用**
Passive DACでの安価な4x100G-400G　Breakoutを実現するため、
サーバ側で使用するNICは100G-QSFP56(50G PAM4)対応を要件とした

「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化⑤

IPv6ベースの社内ネットワークでフラットなL3を構築する

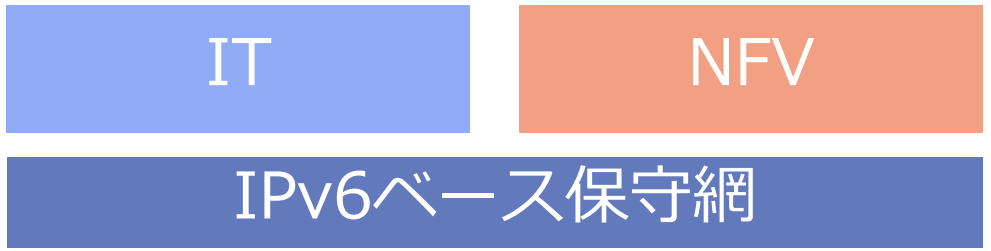
As is



設備文化毎のIPv4ベースの個別NW

- IPアドレスの枯渇
- IPアドレスのバッティング

To be



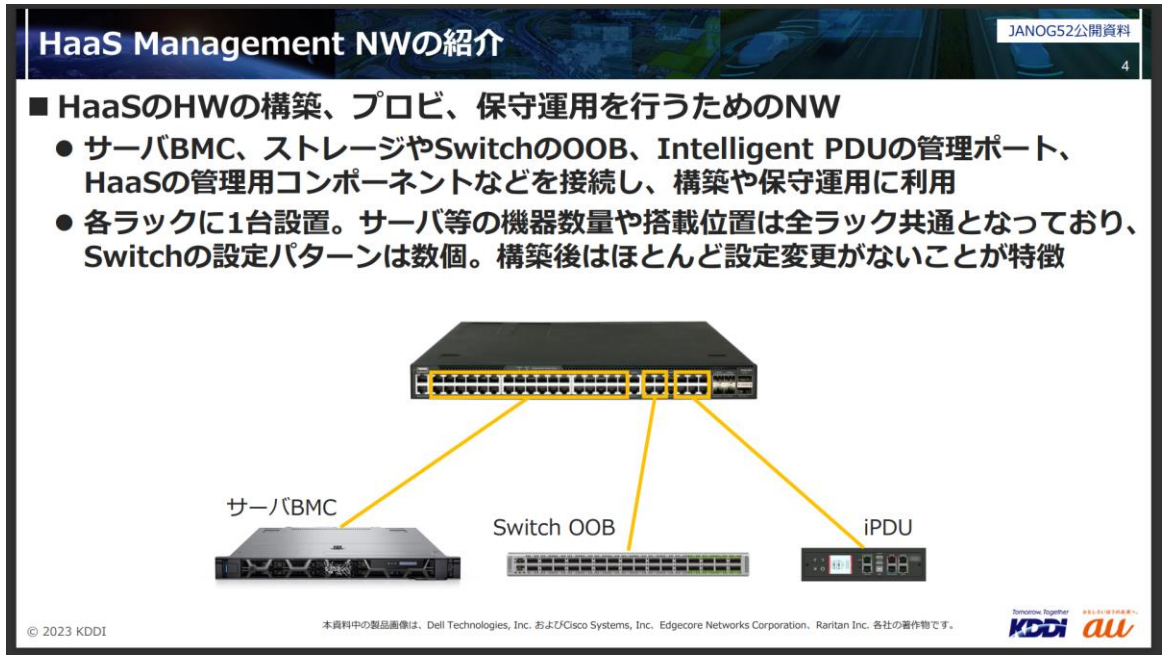
IPv6シングルスタックの統一NW

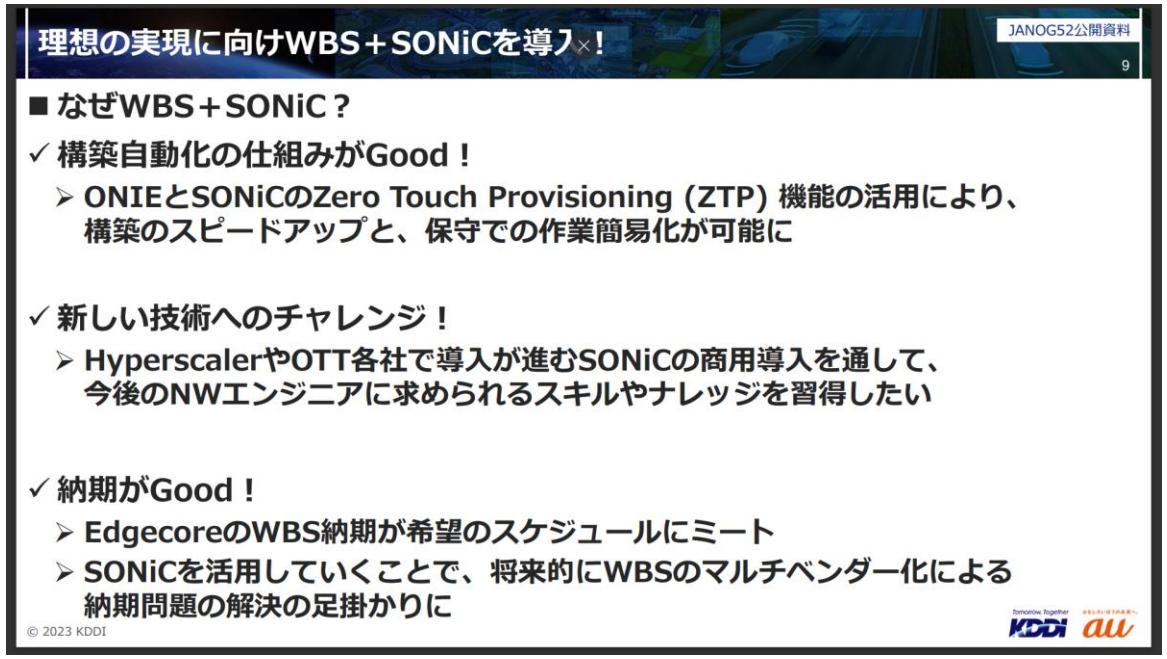
- 潤沢なIPアドレス
- フラットに経路を交換されたNW

- システム間でのサーバ融通を実現するためには現行のつぎはぎのNWでは不足
- IPv6シングルスタックのNWを構築し保守網の一本化を合わせて実施する必要有
- 合わせてDHCP/DNS等の整備

「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化⑥

SONiCを利用したハードウェア管理ネットワーク



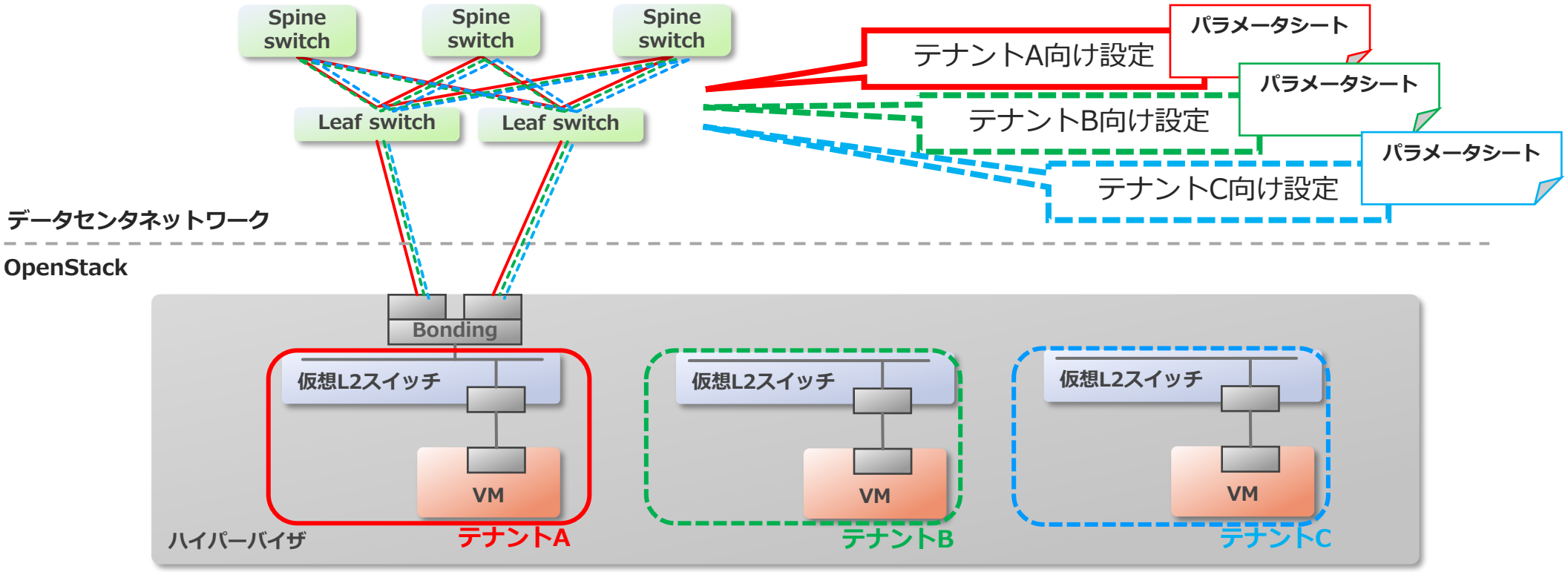


☞ **JANOG52 Meeting @Nagasaki で紹介**

SONiC ZTPでデータセンターネットワークを作った話

「共通仮想化基盤」の構想段階で目指したデータセンタNWの変化⑦

VLANプロビの削減、オーバレイネットワークの活用

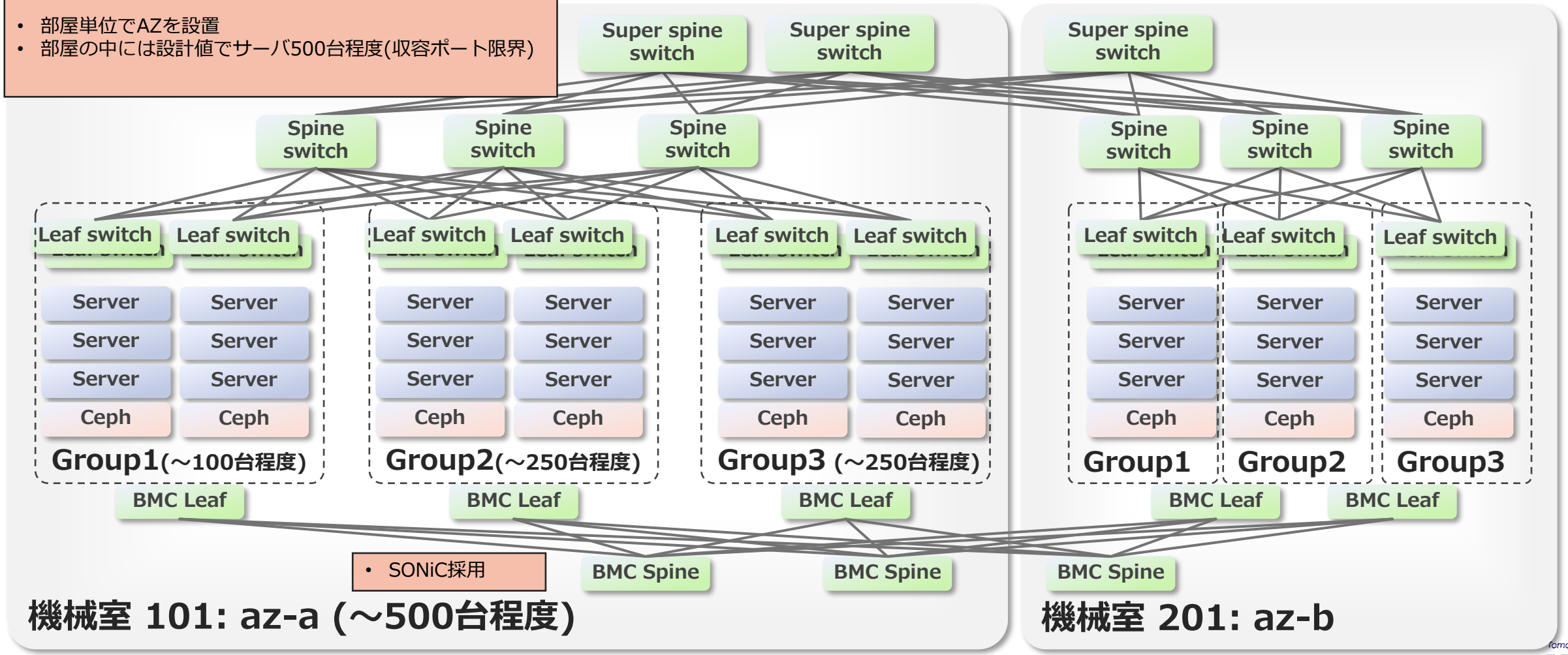


☞ **JANOG55 Meeting @Kyoto で紹介**

NFVプライベートクラウドにおける仮想ルータとBGPによる自律的仮想ネットワーキング

全体的な当初設計のデザイン

- 部屋単位でAZを設置
- 部屋の中には設計値でサーバ500台程度(収容ポート限界)



※絵の記載の数量はイメージであり、実際の数量とは異なります

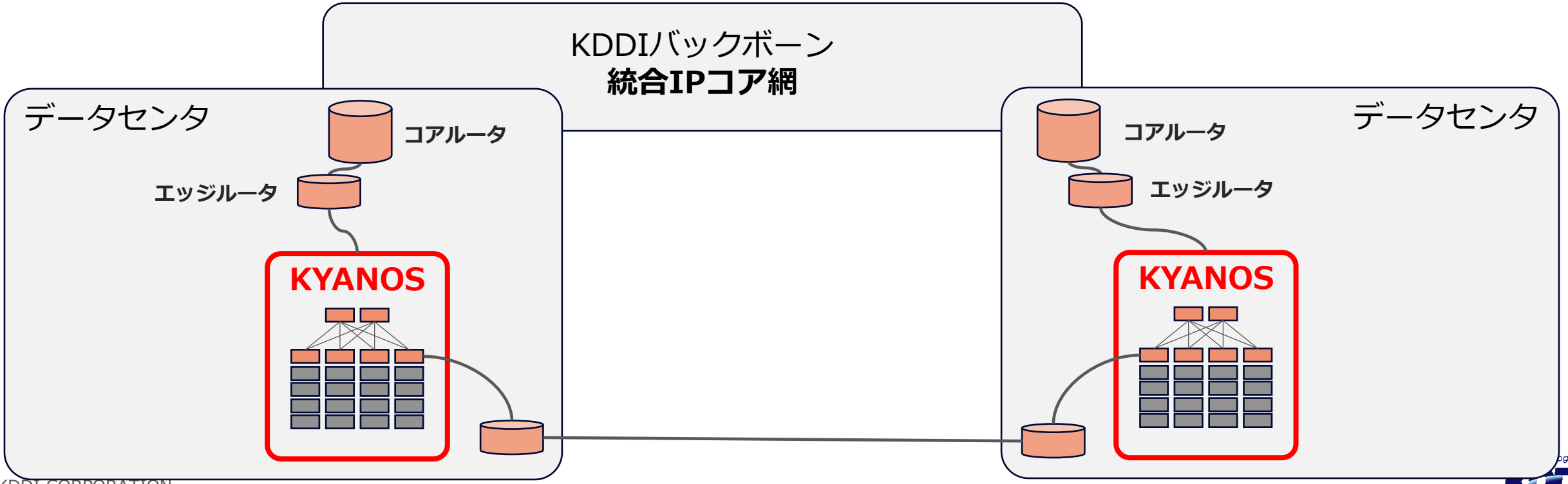
コンセプトの変化（大規模案件の宿命）

途中で変更・追加されたスコープ①：DCI

データセンタ間NWのリアーキテクチャが発生

既存ベンダの延伸ソリューションの導入

- データセンタ間のL2延伸への対応
- コスト効率の良いソリューション

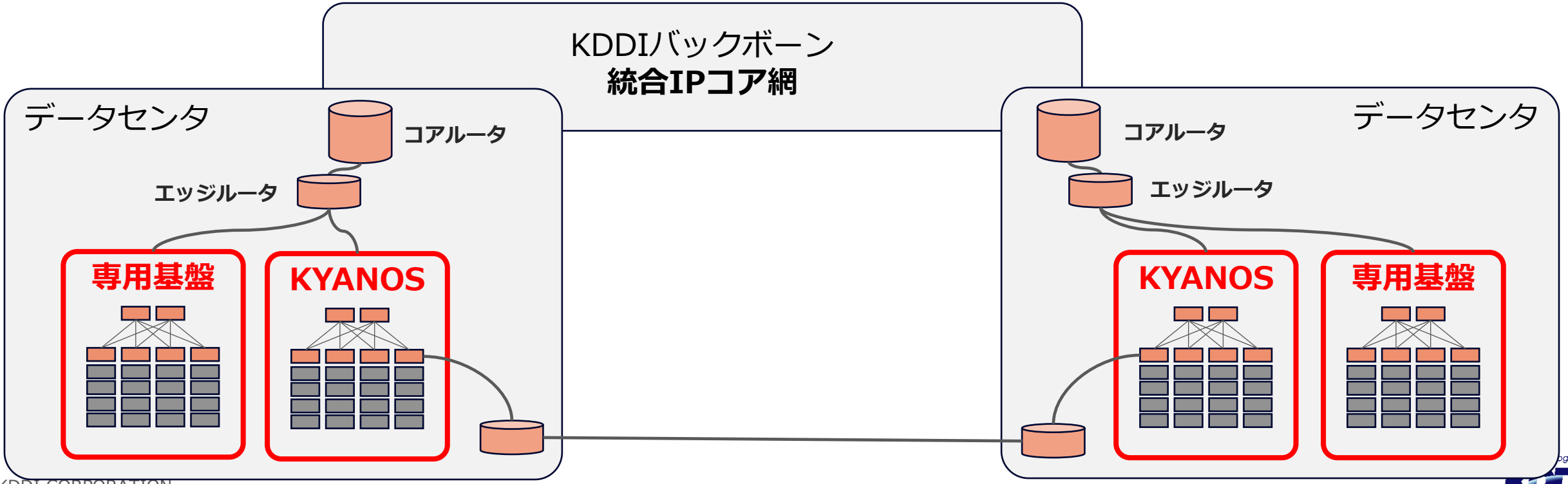


途中で変更・追加されたスコープ②：ベンダ個別基盤提供

特定システム用専用基盤の構築

仮想化に乗れないシステム

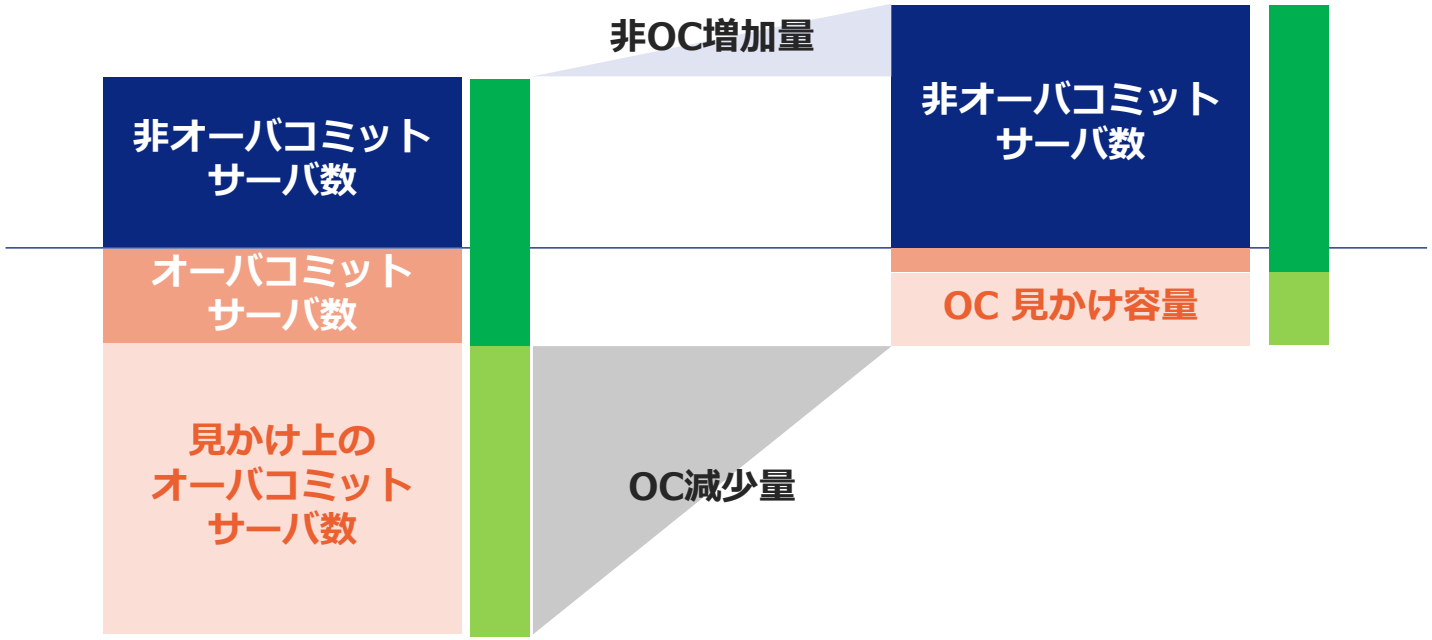
- サーバの（一部）共通化
- ネットワークの設計・ハードウェアの共通化



途中で変更・追加されたスコープ③：設計収容ポリシーの変更

リソース設計ポリシー(オーバコミット適用システム)が当初企画から変更しリソース不足

ラック当たりのキャパシティの変化



- 非オーバコミットは 1.4倍
- オーバコミットは 0.25倍

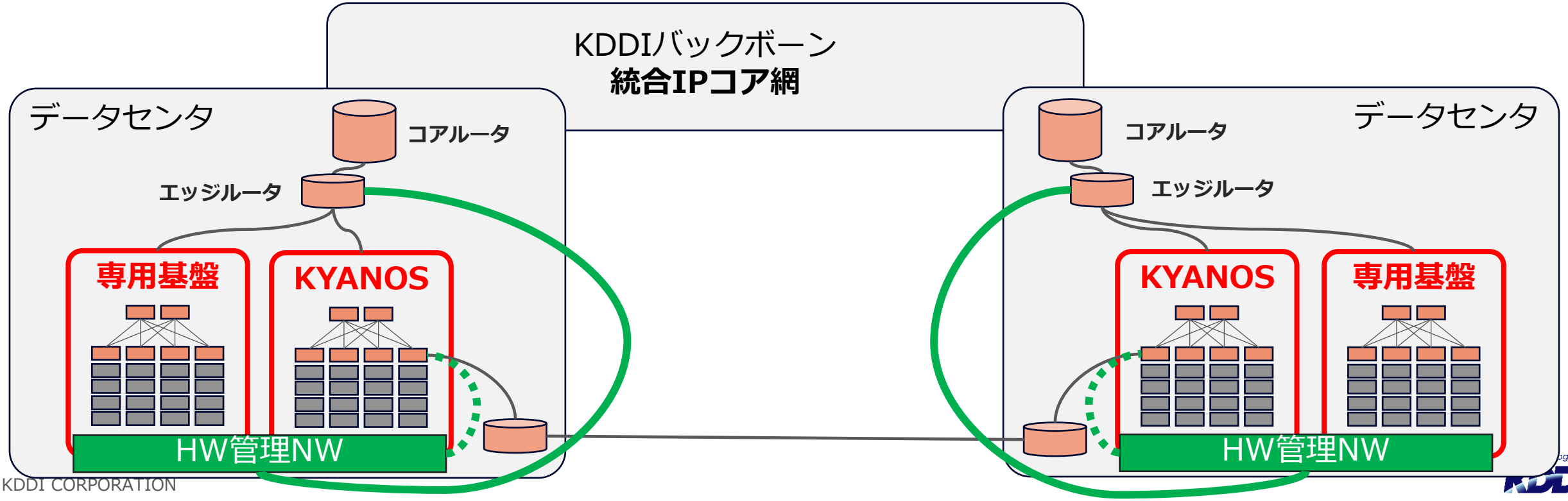
- **全体容量は 0.6倍**

途中で変更・追加されたスコープ④：HW管理NW冗長化ポリシーの変更

最後の手段。どこのレイヤーで分離すべきか…？

最後の手段＝インテリジェントPDUからの電源オフ/OOBアクセス

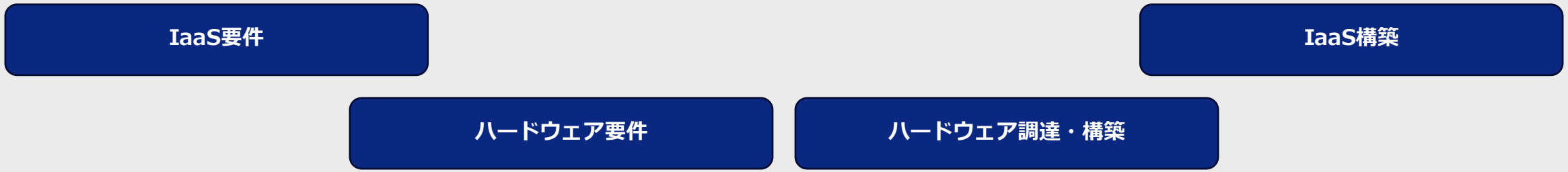
- 本体のNWが死んでいたらどうするのか！？
- バックボーンが死んでたらどうするのか！？！？



＋あいまいなまま進んでいたこと：IaaSの設計に先行してHW調達

リードタイムの長期化等によりHWを先行調達
IaaS/プライベートクラウドは後追いで作ることに

Ideal (要件に最適なサーバ・NW・ストレージを調達する)



Actual (何にでも使いまわせるものを買えばいい/ありものに合わせてIaaSを作ればいい)

実際どうなったのか、どうだったか編

で、実際どうだった？

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌥️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌥️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌥️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

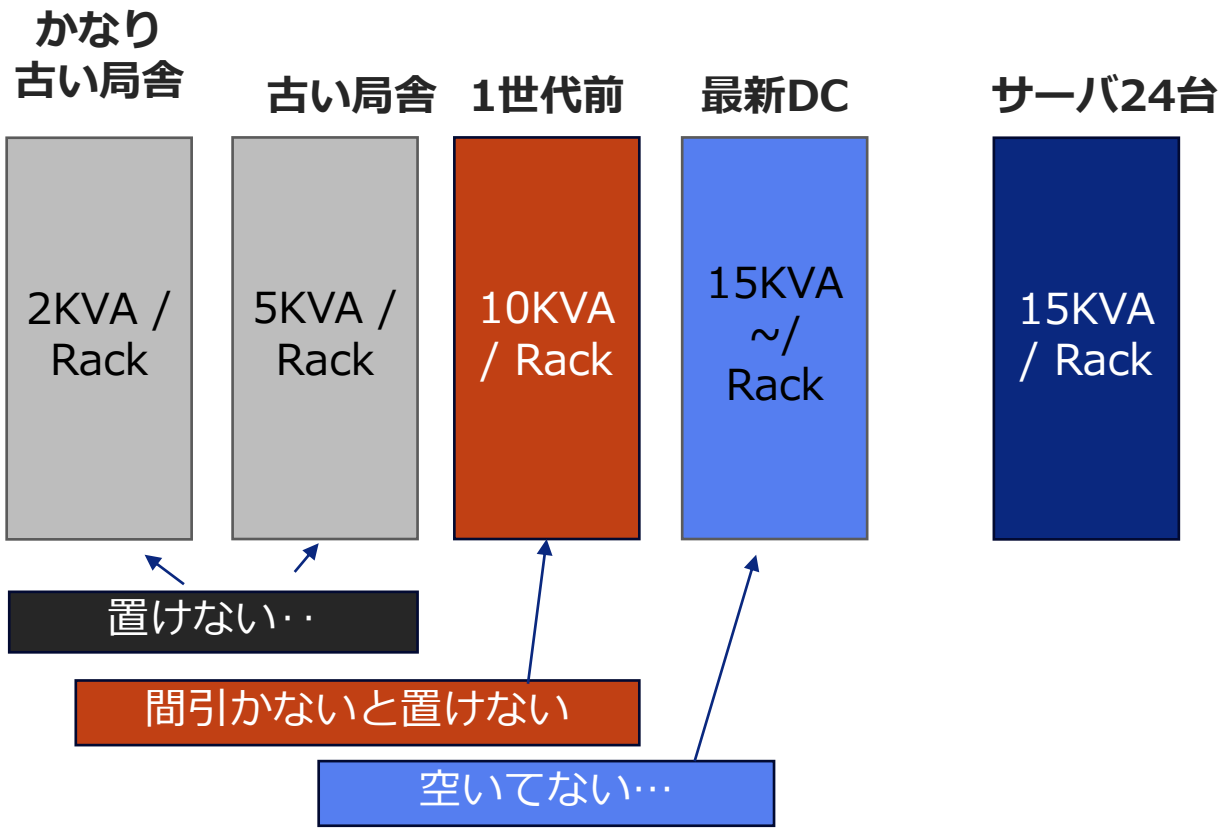
振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌥️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌥️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌥️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

部屋の確保に難航・十分なAZが確保できない①

高負荷ラックを置ける自社DCの空きスペースが少ない



理想

Region A：360台

AZ-A:部屋1	AZ-B:部屋2	AZ-C:部屋3
120/400台	120/400台	120/400台

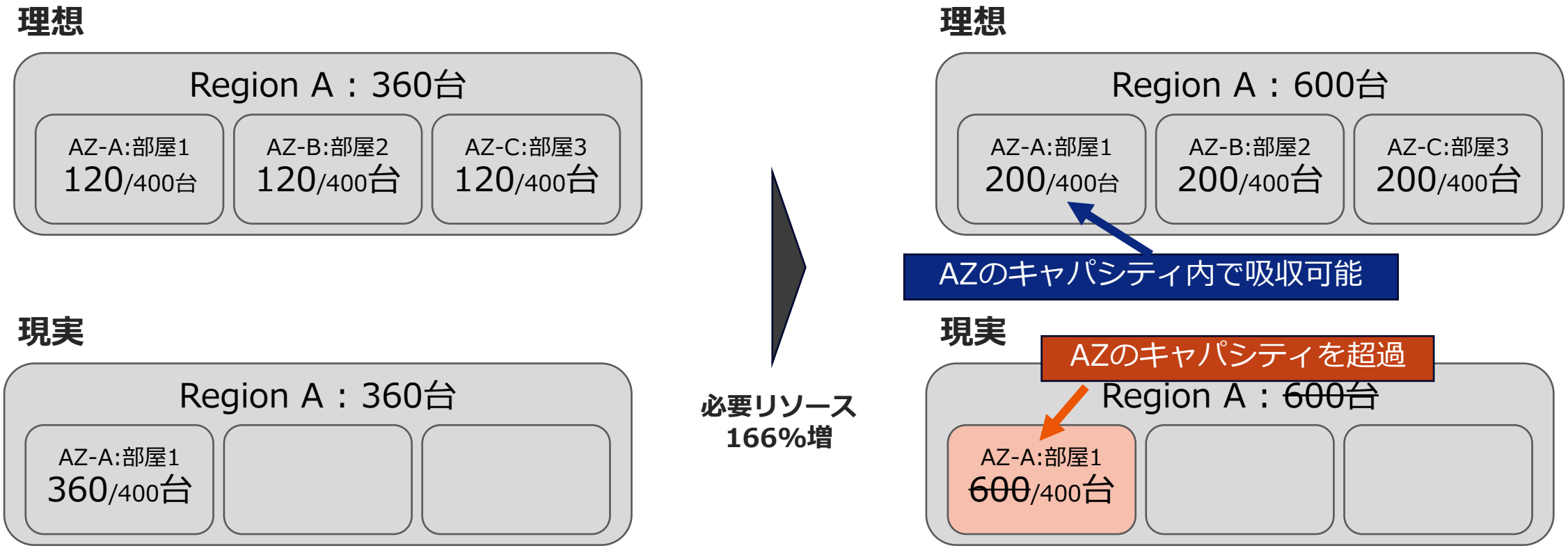
現実

Region A：360台

AZ-A:部屋1		
360/400台		

部屋の確保に難航・十分なAZが確保できない②

リソース設計ポリシー(オーバコミット適用システム)変更によりサーバリソースが不足。
これにより単一AZ偏った構成に対してトドメ。AZのキャパシティが不足。



次で説明する5-StageのスケーラビリティによりAZの増設方法についても課題があり

振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

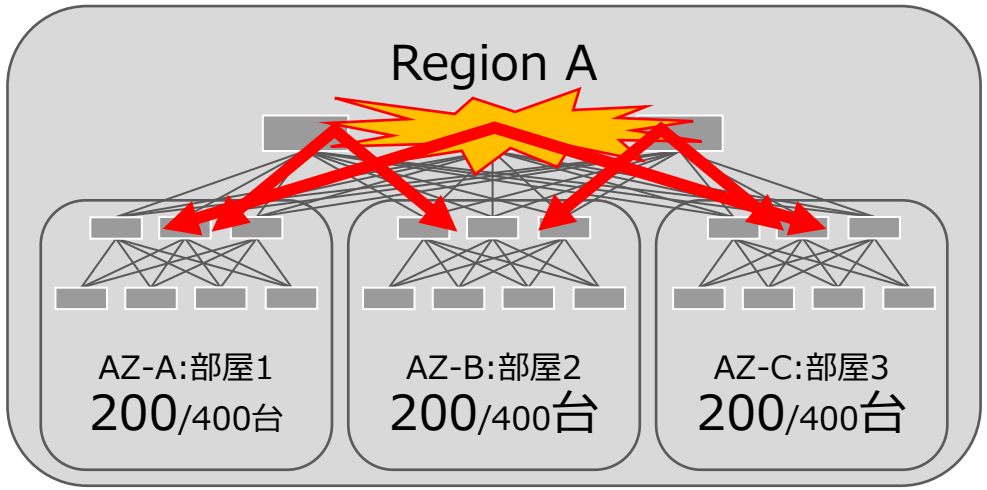
変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌥️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌥️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌥️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

スケーラビリティと障害時の影響範囲に課題①

アベイラビリティゾーンの拡張ができない

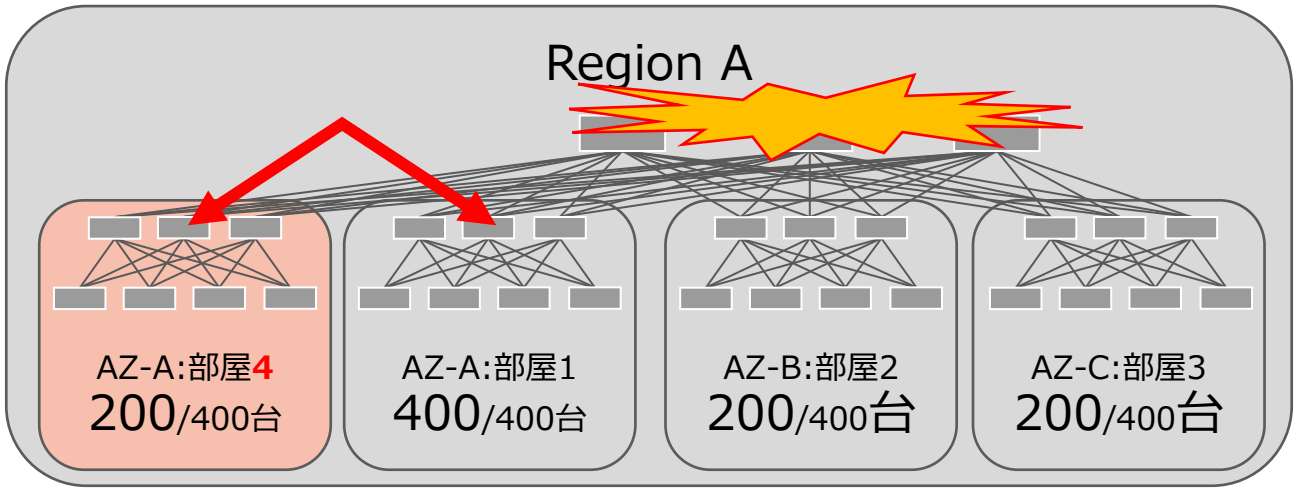
可用性の観点でSuper spineスイッチにAZ**内**通信を通すわけにいかない

当初設計



影響範囲：AZ間通信
AZ間でリソースを共有しない

AZ拡張案



影響範囲：AZ間通信＋AZ-A内通信
AZ間でスイッチのリソースを共有してしまう

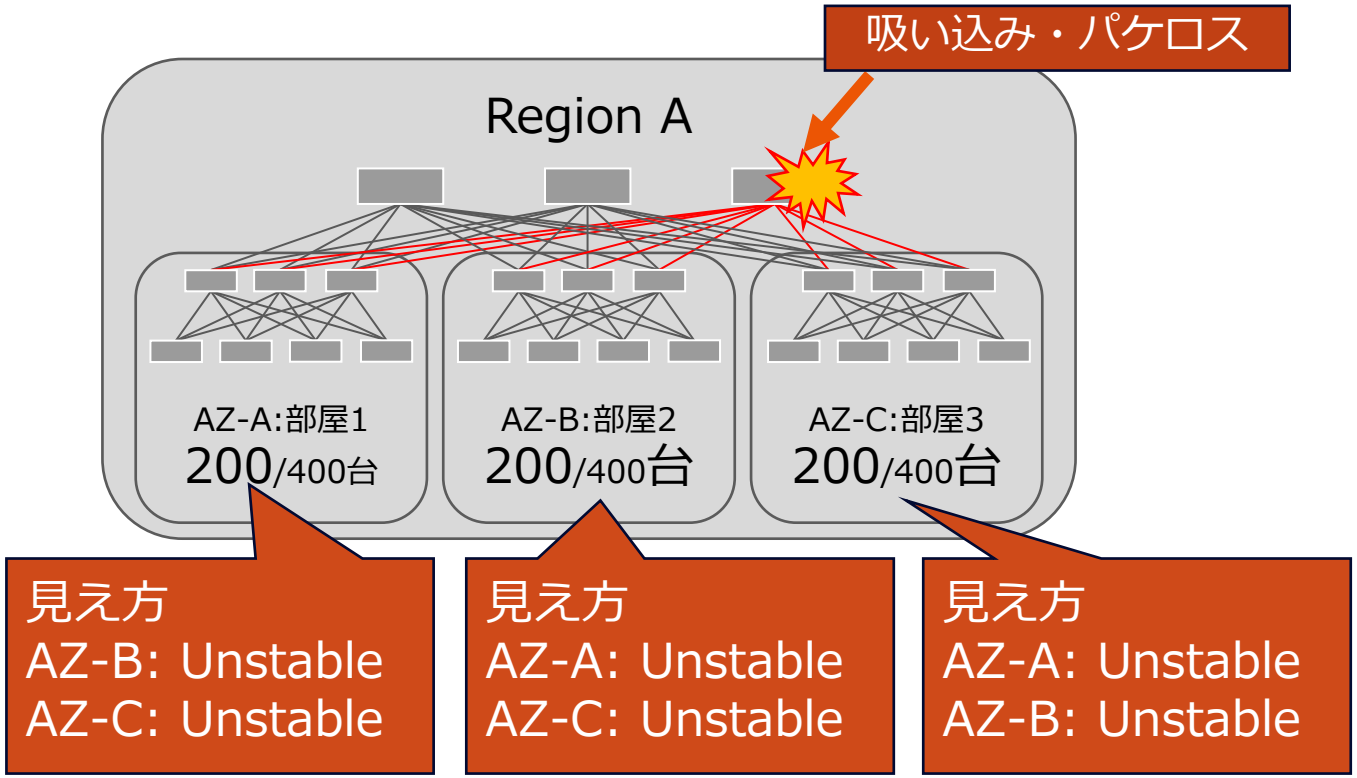
耐障害性要件: すべてのリソースを AZのもの、リージョンなるなもののどちらかに分類したい

AZ間で物理リソースを分離する我々の求める耐障害性要件を満たせない

スケーラビリティと障害時の影響範囲に課題②

アベイラビリティゾーン間通信での影響範囲が大きい

ECMPによりどのAZ間通信にも等しく影響が出てしまい、さらに切り分け困難



ECMPに起因する問題

- スプリットブレイン
- 誤検知の誘発

運用上の苦労

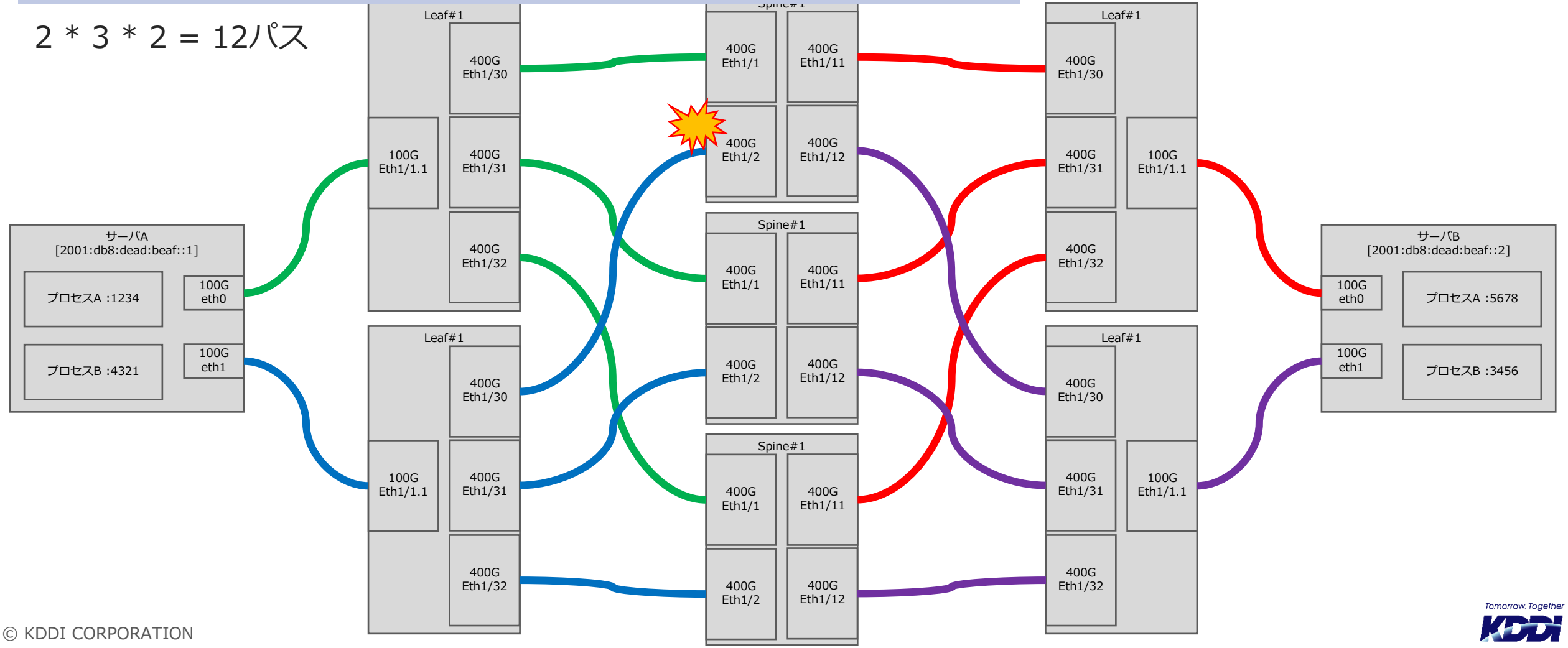
- 障害時の影響範囲や原因の早期特定が困難

補：そもそものClosトポロジでの問題発生時の切り分けのむつかしさ

フロー単位で問題が発生するため原因特定/事前の予測や調査が困難

[2001:db8:dead:beaf::1]:1234 → [2001:db8:dead:beaf::2]:5678

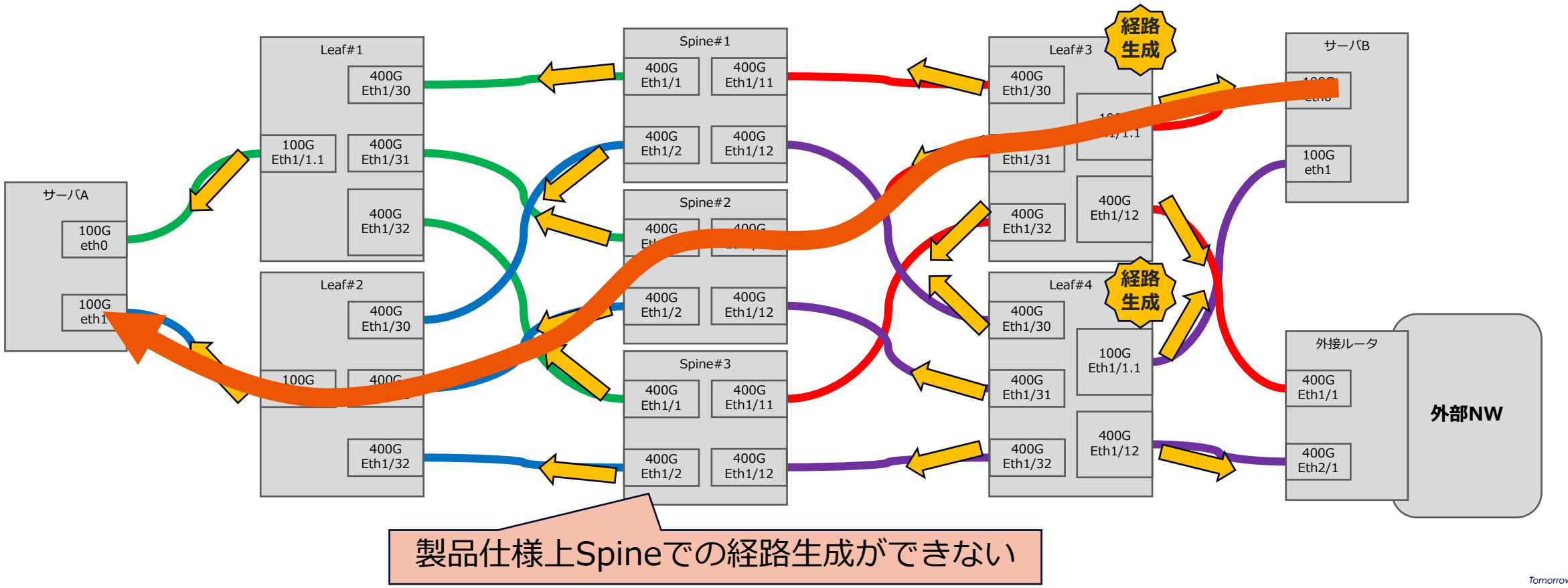
2 * 3 * 2 = 12パス



スケーラビリティと障害時の影響範囲に課題③

経路制御のむつかしさ

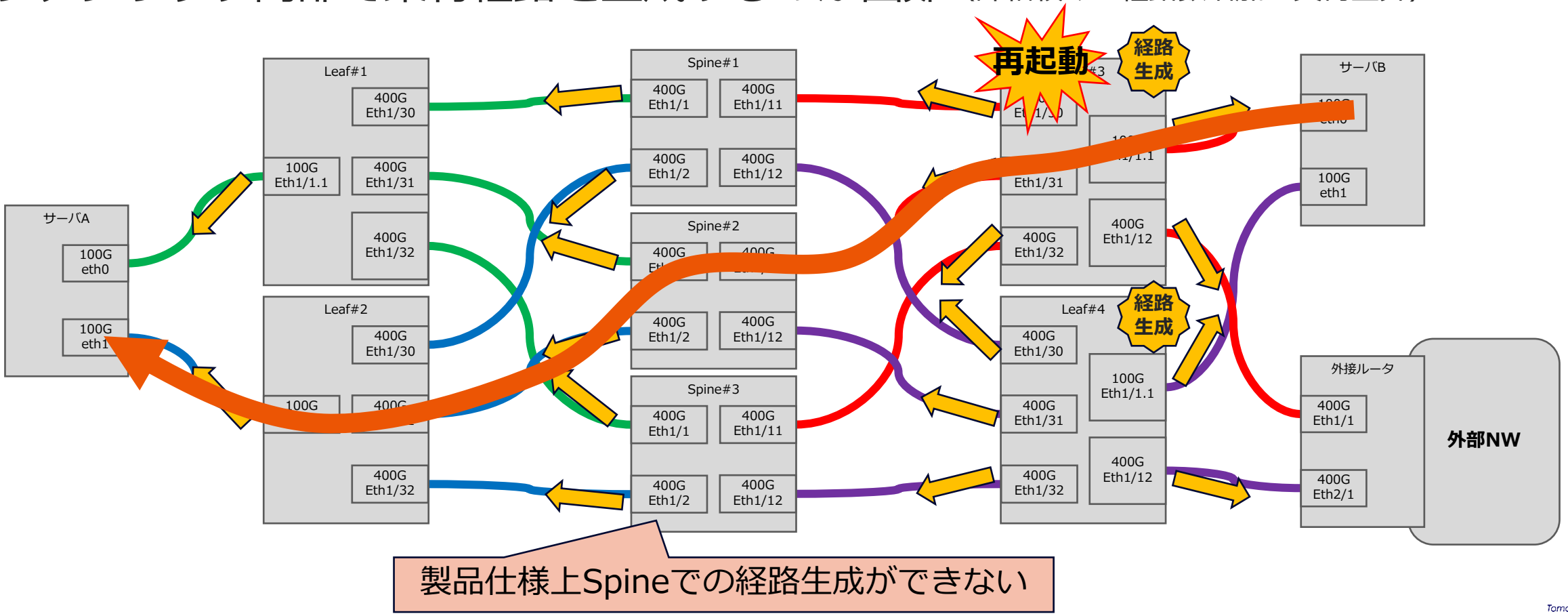
外接ルータと接続するLeafスイッチにサーバを収容してしまったため、
ファブリック内部で集約経路を生成するのが困難（詳細頼み→経路数増加→負荷上昇）



スケーラビリティと障害時の影響範囲に課題③

経路制御のむつかしさ

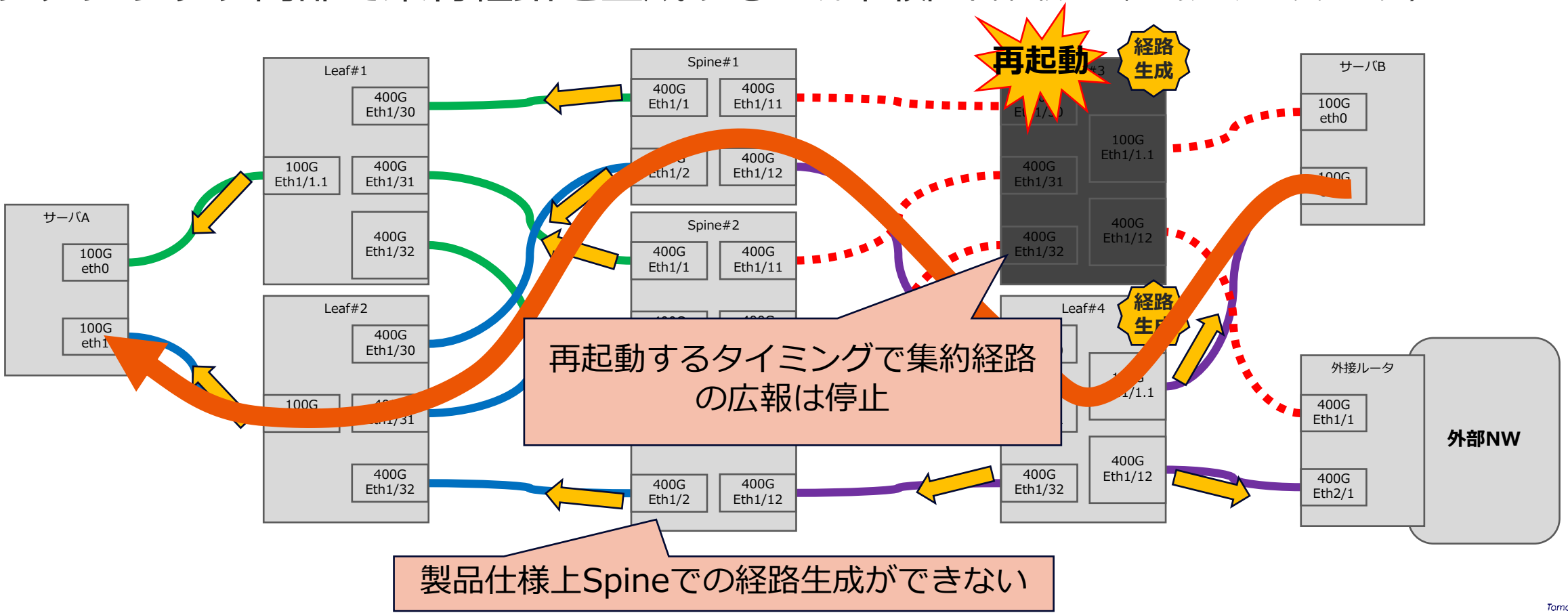
外接ルータと接続するLeafスイッチにサーバを収容してしまったため、
ファブリック内部で集約経路を生成するのが困難（詳細頼み→経路数増加→負荷上昇）



スケーラビリティと障害時の影響範囲に課題③

経路制御のむつかしさ

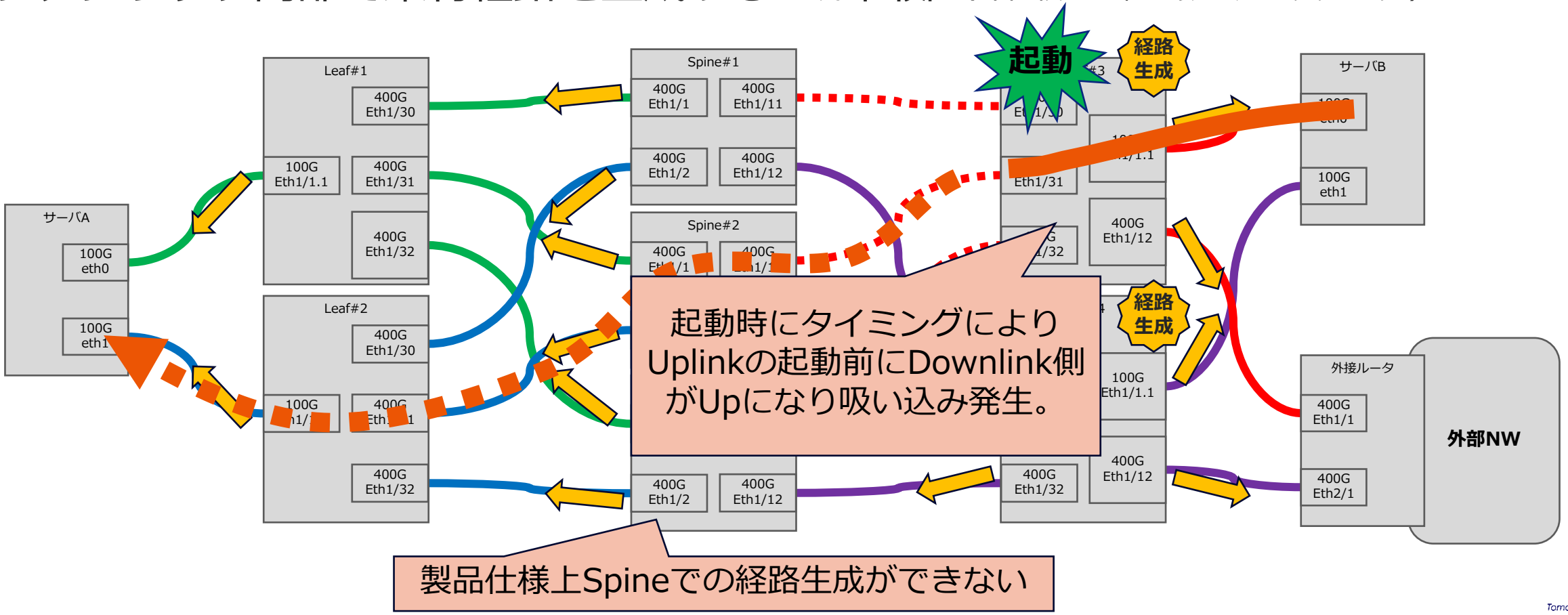
外接ルータと接続するLeafスイッチにサーバを収容してしまったため、
ファブリック内部で集約経路を生成するのが困難（詳細頼み→経路数増加→負荷上昇）



スケーラビリティと障害時の影響範囲に課題③

経路制御のむつかしさ

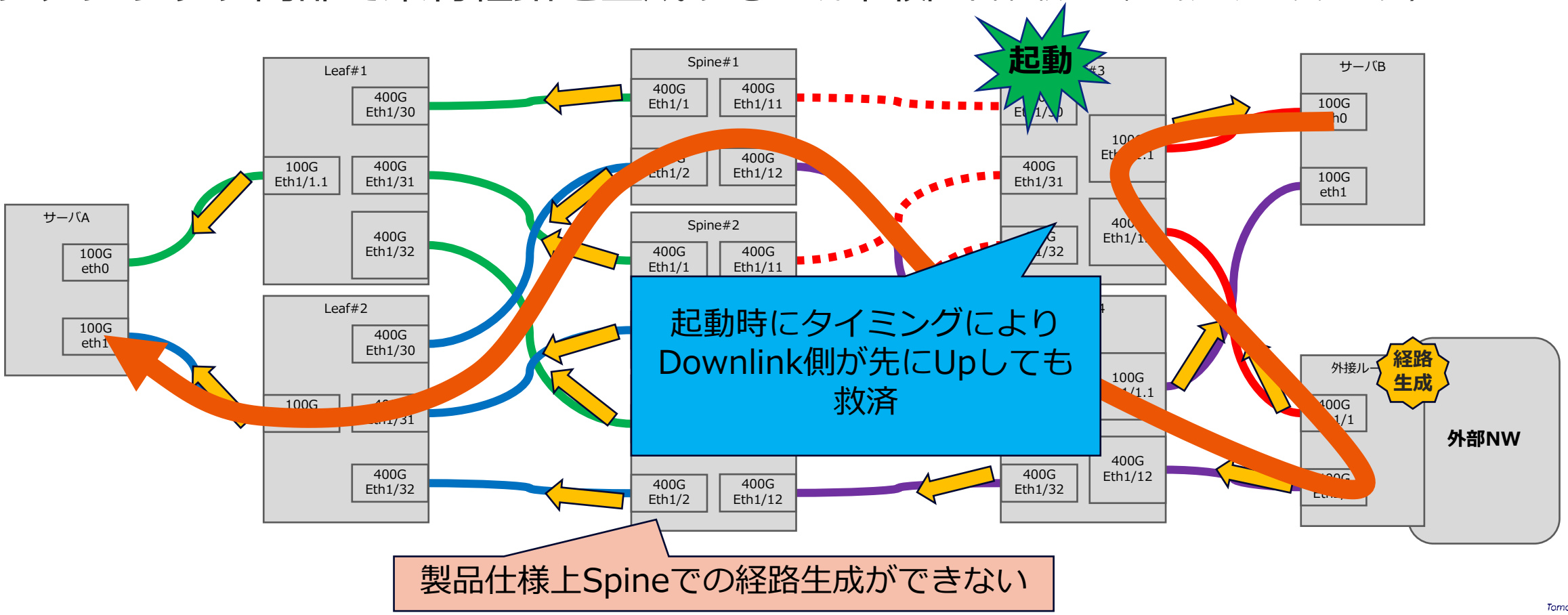
外接ルータと接続するLeafスイッチにサーバを収容してしまったため、
ファブリック内部で集約経路を生成するのが困難（詳細頼み→経路数増加→負荷上昇）



スケーラビリティと障害時の影響範囲に課題③

経路制御のむつかしさ

外接ルータと接続するLeafスイッチにサーバを収容してしまったため、
ファブリック内部で集約経路を生成するのが困難（詳細頼み→経路数増加→負荷上昇）



振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌤️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌤️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌤️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

既存ベンダ製品採用による機能的制約

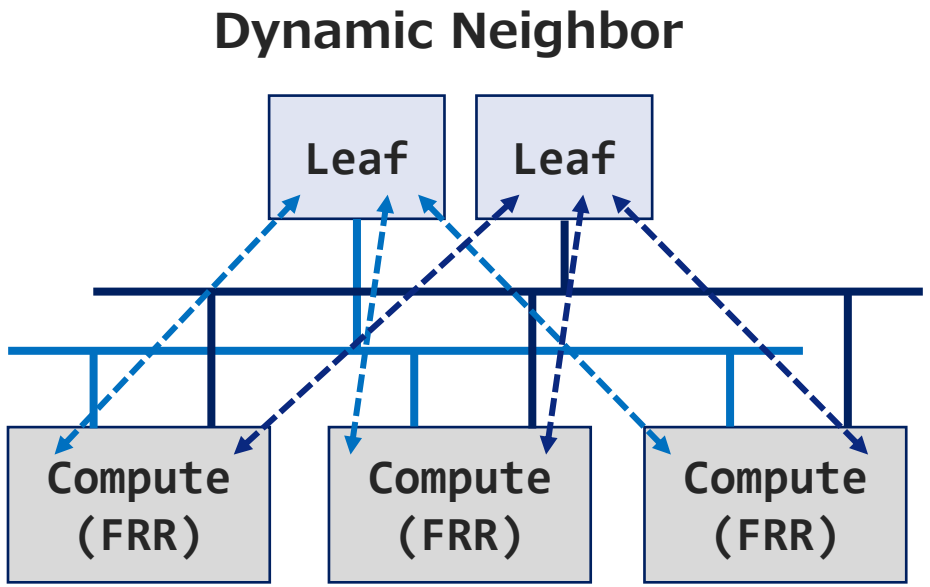
IaaS要件とNW機能の乖離
BGPがL2接続の件/Timer満了まで待たないと切れない

環境

- 多くの台数のコンピュートノードがLeafスイッチとBGPピア接続を行う構成
  (1サイトで1,000台規模、Leaf 1ペアに最大24台)
- 設定量を最小にしたい
- コンピュートノード増設時に都度設定投入したくない

現実

- BGP Unnumbered は Leafスイッチ側未サポートで使用できず
- Leafスイッチ側はDynamic Neighbor設定により設定量を削減し、増設ホストとも自動接続
- バス接続構成によって、I/F障害時もBGPピアが落ちない
  (ComputeサーバのNIC障害なら断検知可能な場合も)
- しかし、Leafスイッチ側で Dynamic Neighbor設定とBFDが同時に設定できないことが判明し、障害検出はタイマで実施

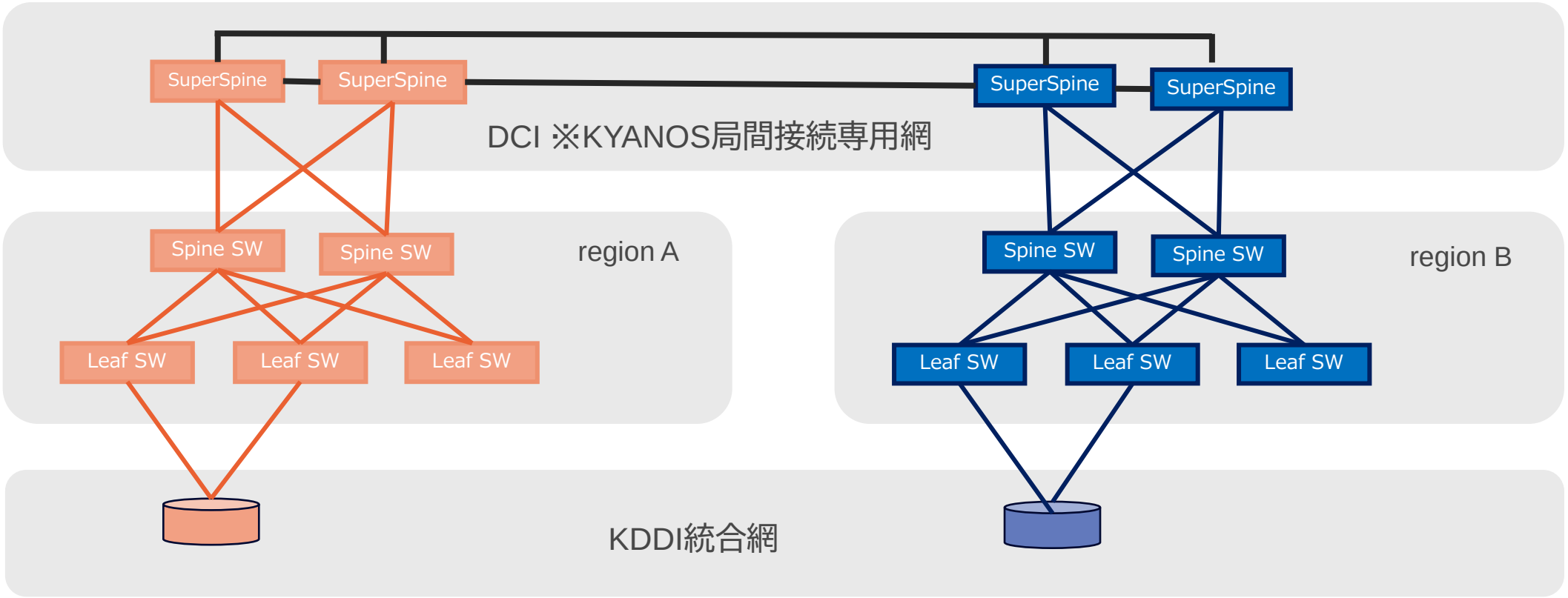


マルチサイトソリューションの破綻①

KDDI統合網とは別にDC間専用回線にてregion間を接続

・region間はKDDI統合網ではなくDCIにて通信
・某ベンダのVXLAN延伸ソリューションを採用

共通基盤局間接続イメージ図



マルチサイトソリューションの破綻②

メリット

- L2でregion間を延伸可能
- プロビジョニングが迅速
  - コントローラからワンタッチでSite間通信を延伸可能

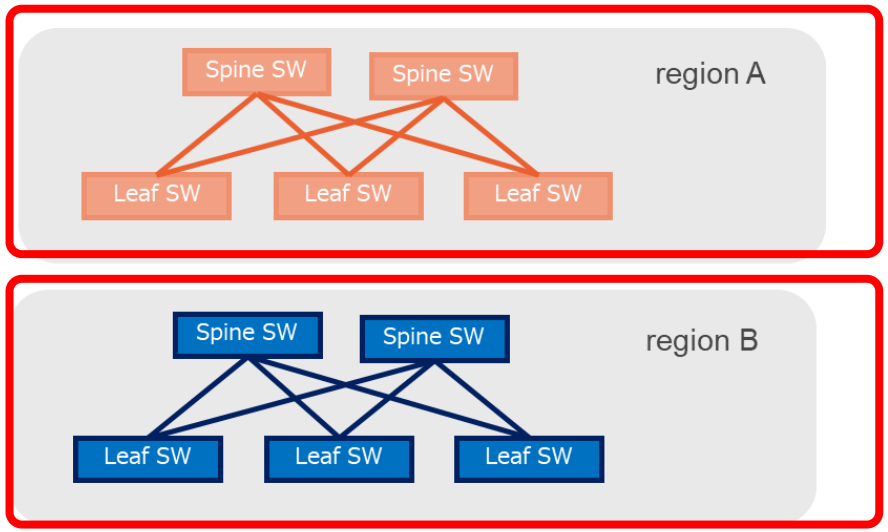
デメリット

- 経路制御が複雑化
- 採用したソリューション特有のスケーラビリティが存在

マルチサイトソリューションの破綻③

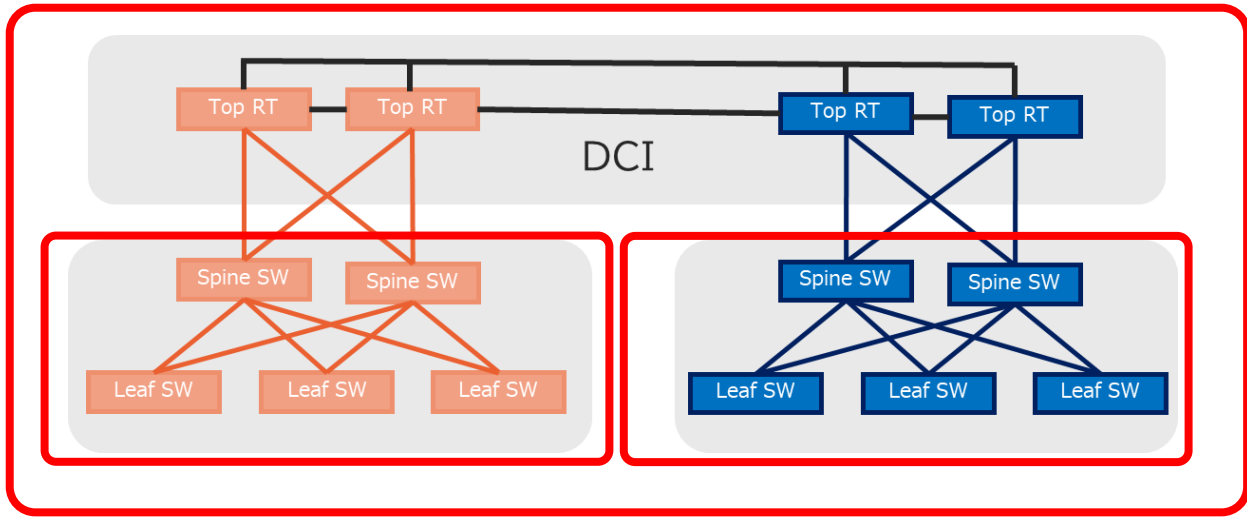
スイッチ数の総計、収容可能テナント数、VRF数などの上限値が存在

DCIなし



設定上限値を意識すべき範囲

DCIあり



設定上限値を意識すべき範囲

各regionごとにscalabilityを気にしていればいい

←に加えて全region合計数にもscalabilityがある

DCI特有のscalabilityの一部が致命的で、DCIは解体

マルチサイトソリューションの破綻④

メリット

- L2でregion間を延伸可能
- プロビジョニングが迅速
  - コントローラからワンタッチでSite間通信を延伸可能

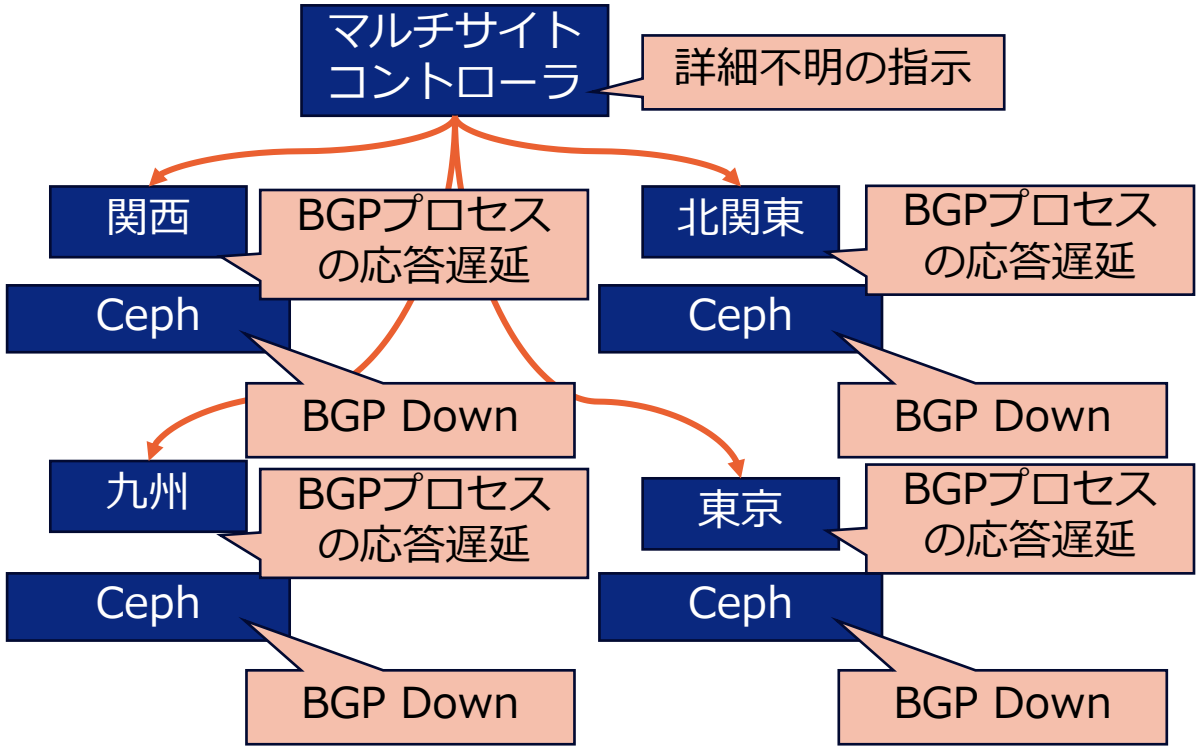
導入後、上記をテナントが求めていないことが判明

重点議論のポイント①

ベンダロックイン構成

マルチサイトソリューションに起因する作業時に全サイトでBGPセッションが同時にフラッピングが発生する問題に直面

どうしてそうなるのかの検討もつかない、内部は見えるけど教えてもらえない



限定的なコマンドから見える結果
- Non-root
- 限定的なコマンドライン(打てないコマンド)
- /proc/…から見える状況

何となく推察される結果…
- CPU使用率
- CPU Pinningはされてない・・
- スケジューラはSCHED_OTHER・・

開示されない内部実装
- 「わからない」「開示できない」

振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌤️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌤️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌤️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

補足 DACケーブルについて

DACの構造

SFP部分とケーブル部分が一体

一般的なDACのメリット

安い

・光ファイバと同等の性能を持つにも関わらず、コストが1/2~1/5程度

耐久性が高い

・光モジュールとケーブルがシームレスな接続方式のため、ほこりなどに強い

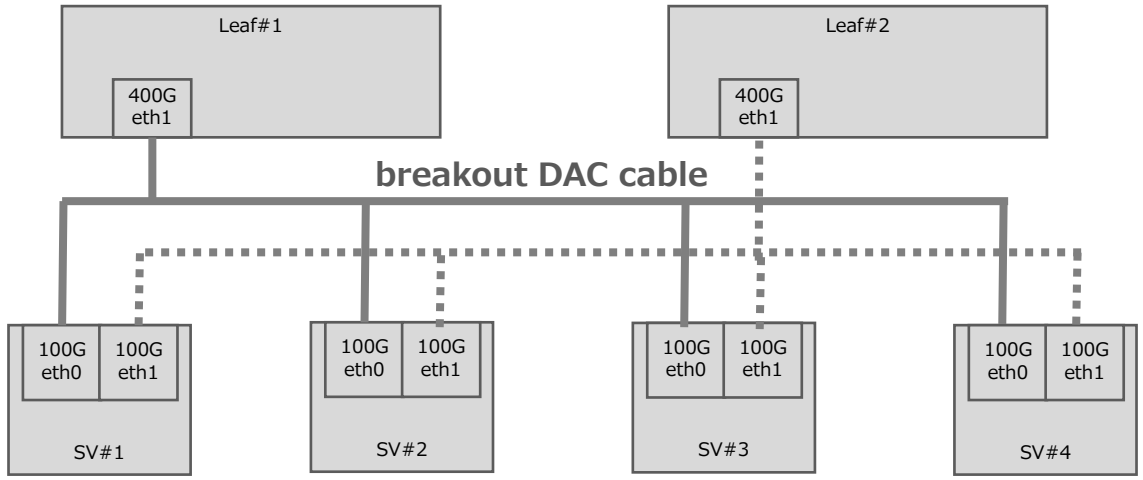
DACの運用で苦労した話

Input errorが多発するcableあり

- Leaf~SV間を400G～100G×4 Break out DAC cableで接続
- メーカー検証で「DAC+4分岐」のパターンのみerrorが発生することが分かっている
- 電磁干渉(EMI)が原因と考察されているが、原因の特定には至っていない

①銅線(DAC)は電気信号から電磁場を生成＆電磁干渉の影響を受ける
②break out cableはストレートcableと比較して密集して配置されるため、EMIの影響を受けやすい

Leaf-SV間構成



Leaf-SV間を400G～100G×4 breakout cableで接続

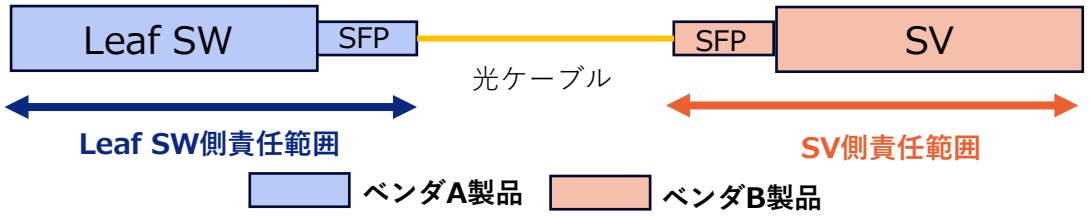
エラー発生パターン

ケーブル種別	光ケーブル		DAC	
分岐有無	ストレート	4分岐	ストレート	4分岐
信号媒体	光	光	電気	電気
事象発生	無	無	無	有

DAC+4分岐のパターンのみerror発生

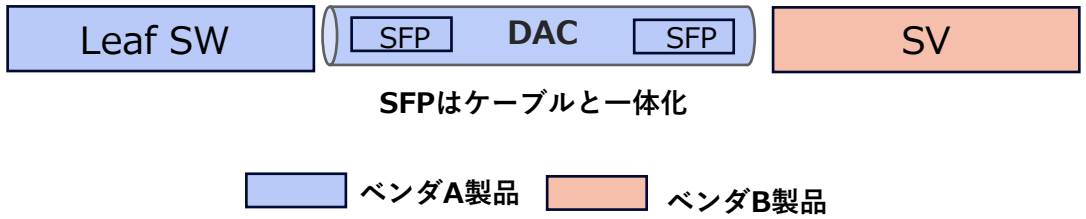
DACの運用で苦労した話①（責任分界点）

一般的なSFP/光構成の場合



機器と搭載するSFPの製造ベンダを揃えることが可能
⇒SW側/SV側ともにベンダサポート対象

DAC構成の場合



機器~Optics区間で製造ベンダ違いが発生
⇒大抵の場合、非サポート
⇒不具合発生リスクが上昇

※実際にSVベンダ検証にてSV非サポートDAC利用時にinputエラーが多発する報告あり

DACのコンセプト(ケーブル/SFP一体)的に
「不具合を引きやすい」＆「不具合を引いたときのダメージが大きくなりやすい」

DACの運用が難しい話②

Leaf SW Port利用イメージ図

Eth1/33		Eth1/1	Eth1/3	Eth1/5	Eth1/7	Eth1/9	Eth1/11	Eth1/13	Eth1/15	Eth1/17	Eth1/19	Eth1/21	Eth1/23
		使用不可	DAC用	使用不可	DAC用	使用不可	DAC用	使用不可	DAC用	Spine向け			
Eth1/34		Eth1/2	Eth1/4	Eth1/6	Eth1/8	Eth1/10	Eth1/12	Eth1/14	Eth1/16	Eth1/18	Eth1/20	Eth1/22	Eth1/24
		DAC用	使用不可	DAC用	使用不可	DAC用	使用不可	DAC用	使用不可	Spine向け			

初期導入時に選択したDACケーブルが太すぎて SW全PortにDACをさせない

図のように使用禁止Portあり状態で運用（SWのPort無駄になるが本当に安い？）

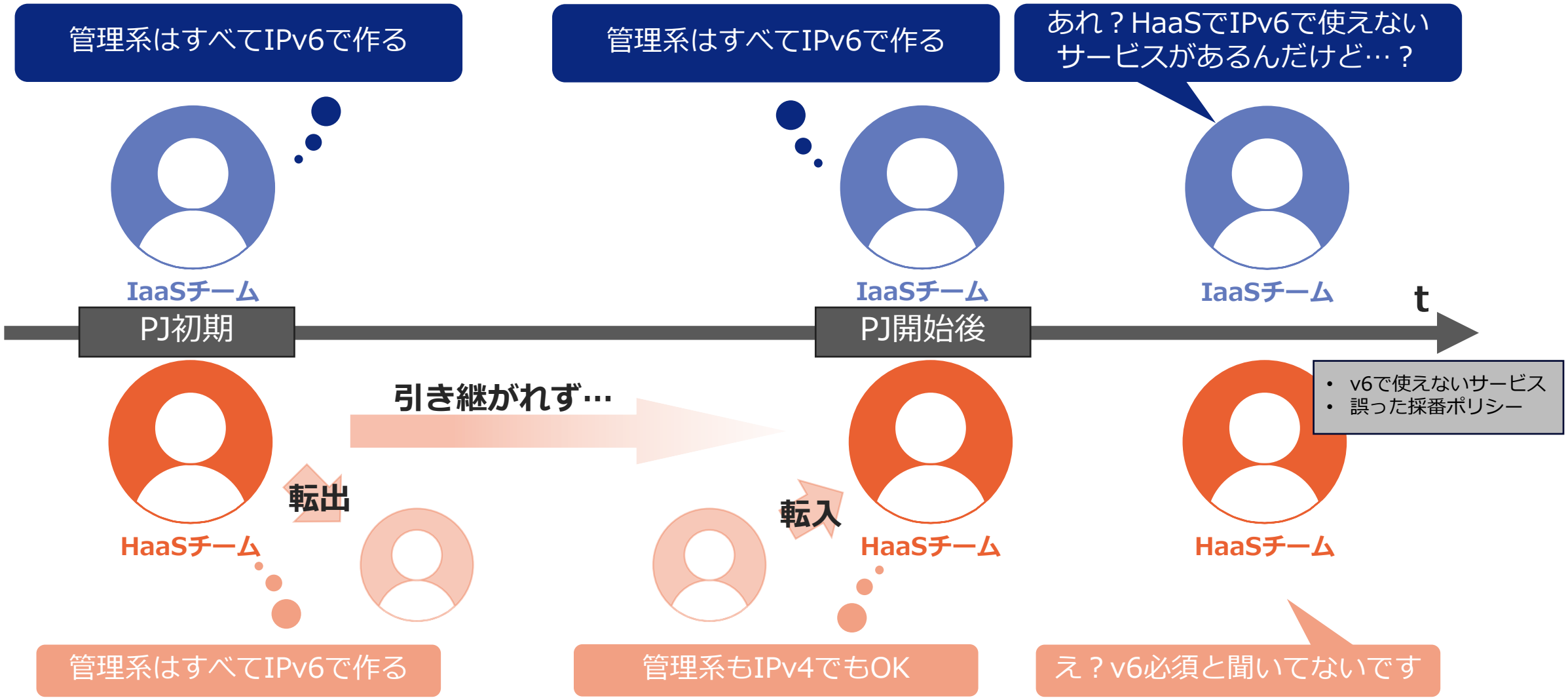
振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌥️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌥️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌥️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

徹底されないシングルスタック利用

初期検討メンバが入れ替わるなどで部内でも認識齟齬が発生…



振り返りサマリー（再掲）

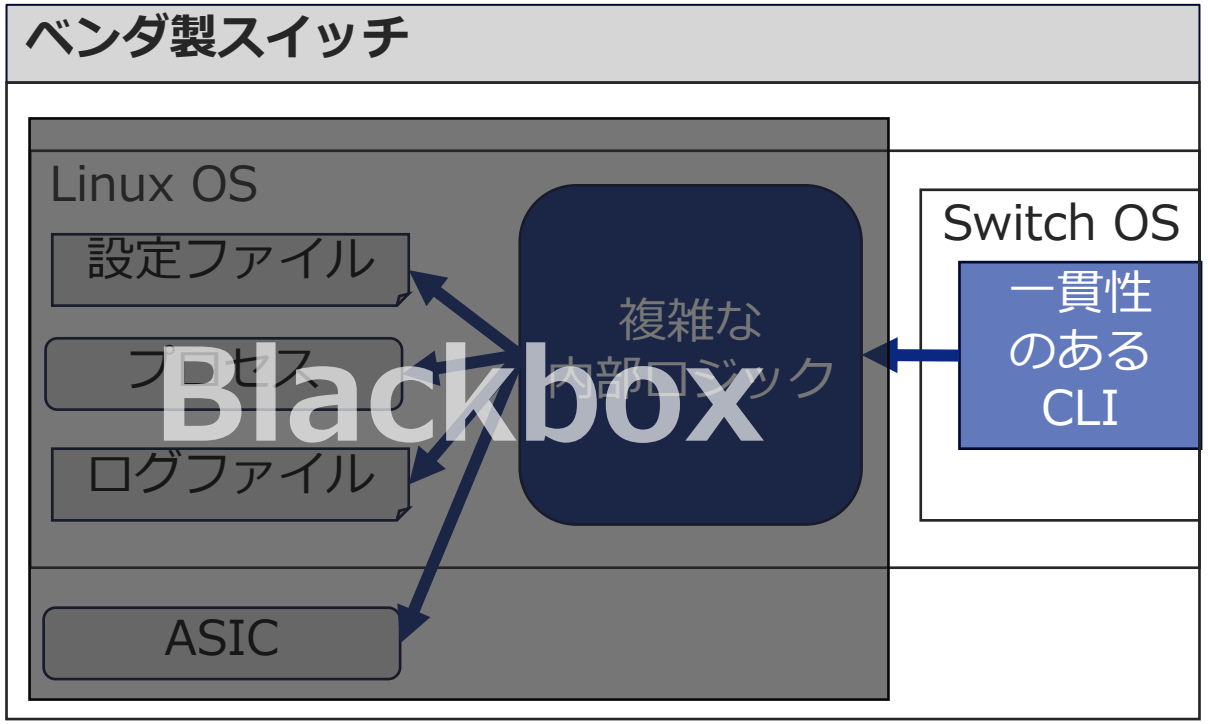
失敗の類も含めて課題は多い…

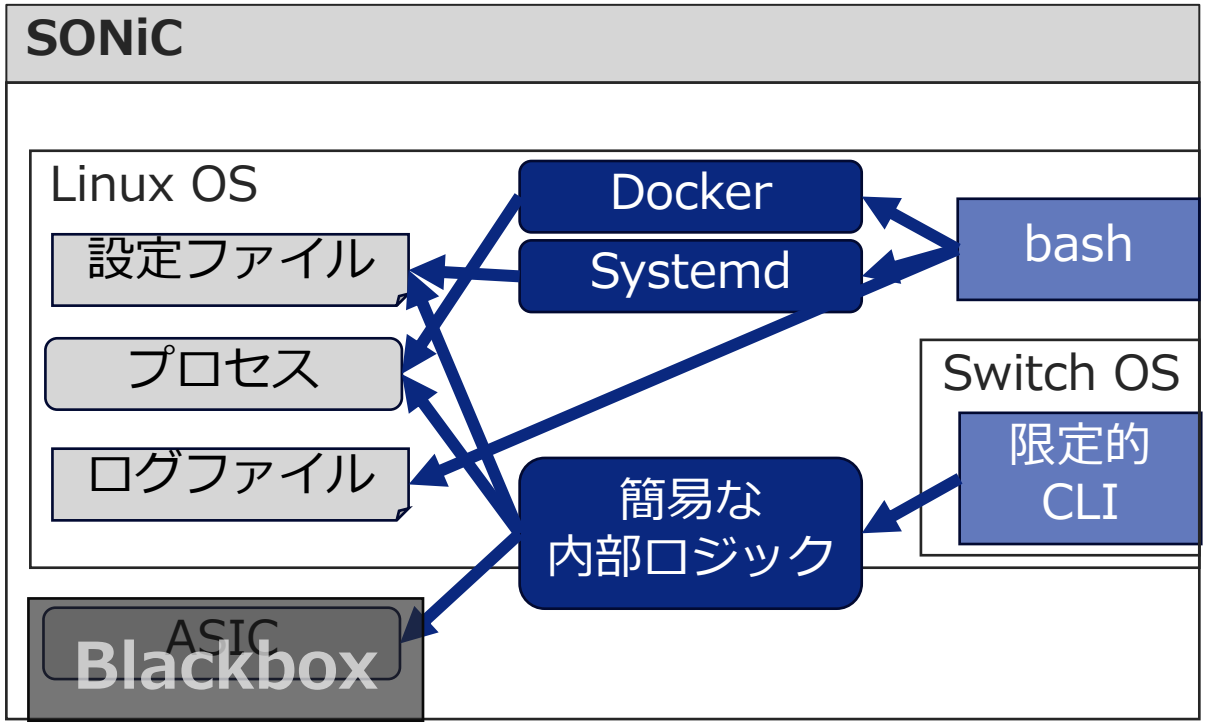
変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌥️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌥️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌥️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

SONiC導入するも安定化に苦労‥その後の広がりが…①

「ネットワークエンジニア」にとってトラブルシュートに苦労

「ベンダ製スイッチ」のやり方ではSI+トラブルシュート共にうまくいかない。
Linuxサーバ+OSSソフトウェアとしての付き合い方が必要。



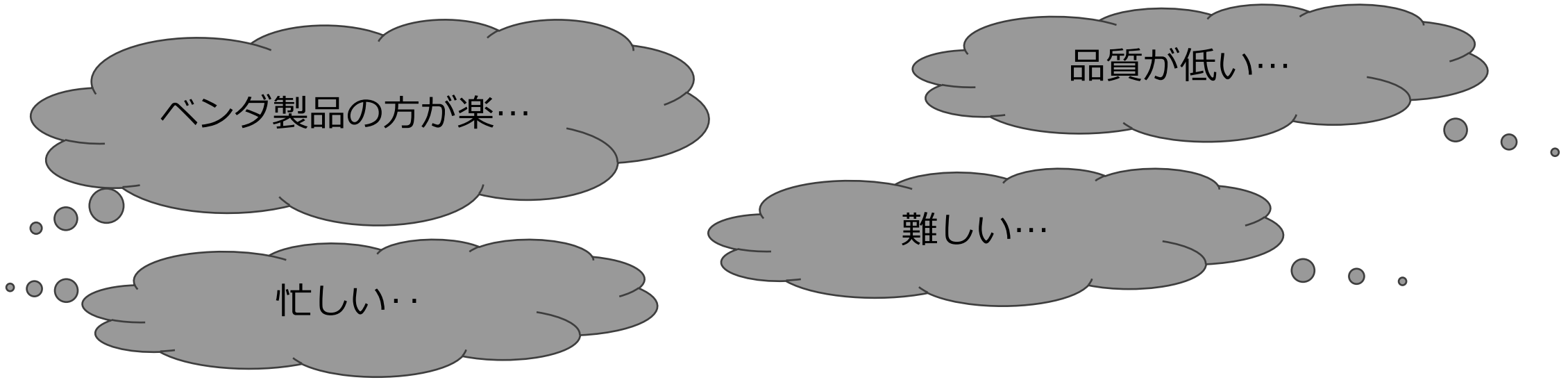


”誰もがこの働き方を気に入るわけじゃない”

SONiC導入するも安定化に苦労‥その後の広がりが…②

”誰もがこの働き方を気に入るわけじゃない”

メンバの意識として 積極的に活用していこう！とはなっていない状況。



更なる活用はスキルセットやエンジニアリングの文化や風土にも大きくかかわる

振り返りサマリー（再掲）

失敗の類も含めて課題は多い…

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌤️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌤️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌤️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

オーバレイネットワークの利用はうまくいった

仮想ルータ等を活用することで大幅に稼働減

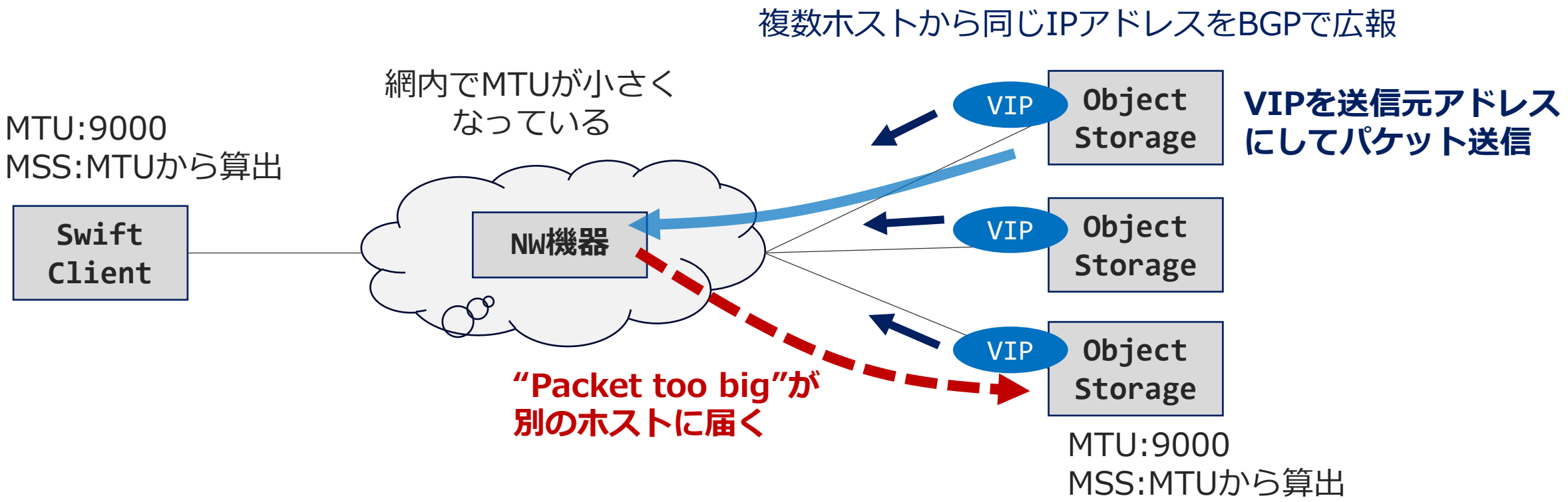
空いた稼働は…

増えた専用基盤の工数に

☞ JANOG55 Meeting @Kyoto で紹介

NFVプライベートクラウドにおける仮想ルータとBGPによる自律的仮想ネットワーキング

IP Anycast + マルチパス利用時の考慮事項



- 負荷分散と冗長構成を目的に、サービスのエンドポイントとなるIPアドレスを複数のホストに設定して、それぞれからBGPで広報
- 1台のホストから、送信元アドレスをエンドポイントのIPアドレスに設定して通信を開始したところ、途中のNW機器からの”Packet Too Big”が別のホストに届いてしまう事象に遭遇
- ホスト側のMTU値を中継網に合わせることで解決

これからどうしよう編

振り返りサマリー（再掲）

アーキテクチャや製品選定の見直しを検討

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌤️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌤️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌤️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

アーキテクチャ見直し

製品見直し

アーキテクチャの停滞

課題が課題として認識されない

- 多忙・人員不足 ➡ 積極的な変化のモチベがわきにくい
- 木こりのジレンマ・現状維持バイアス

短期的目線

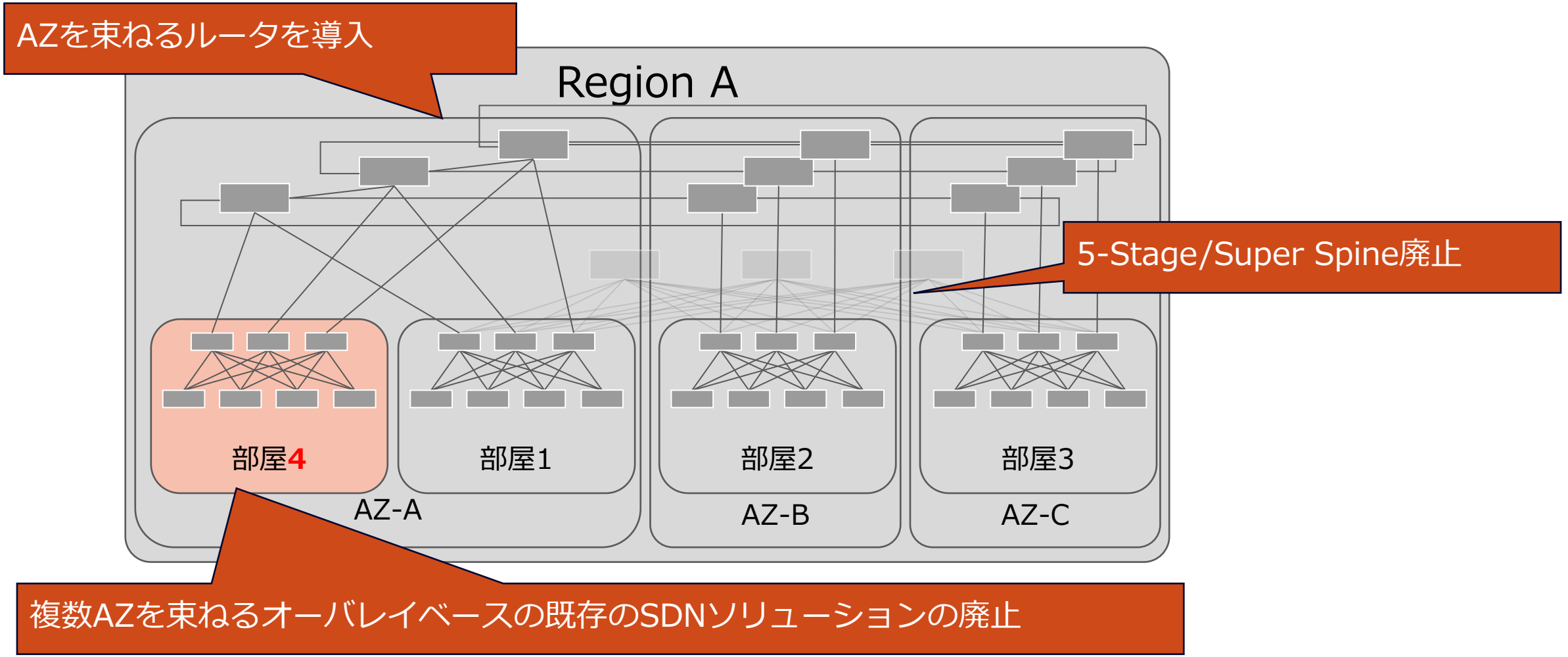
現状維持バイアス

コスト削減意識

➡ **Re-architectureの阻害**

アーキテクチャ変更を検討中

AZの接続方法を5-StageからAZ-awareなルータに変更
機器台数は増えてしまう点が課題



DAC＋Breakout・・・辞めます

光にします。HWコスト・・・増😿

とはいえ、今のままだと運用コストが高すぎる

部材交換・予備品 ＋ **交換対応の稼働**

- エラー検知
- 自動閉塞
- 交換判断
- 交換日程調整
- VMのライブマイグレ
- 物理交換
- 正常性確認組込み

まとめ

まとめ

チャレンジには失敗もつきもの…

- やってみて初めて分かること、検討が不足していたこと、
  様々な事情から（現場レベルでは）どうしようもなかったこと…
- 失敗から学ぶことも多いのでこれを糧によりよりNWへ向かってチャレンジ中です！

変化	結果サマリ
①DCNWの設置ポリシー見直し (エリアから部屋)	🌤️ 部屋の確保に難航・十分なAZが確保できない
②5stage化（3Stageから5Stage)	☔ スケーラビリティと障害時の影響範囲に課題
③データセンタネットワーク製品の流動性	☔ 既存ベンダ製品採用による機能的制約、 マルチサイトソリューションの破綻。 結果的ロックイン
④DACケーブル+Break-out採用	⚡ 品質問題の発生
⑤IPv6シングルスタック	🌤️ 徹底できないポリシ・・・
⑥HW管理NWでのSONiC利用	🌤️ 安定化に苦労‥その後の広がりが…
⑦オーバレイネットワークの活用	🌞

議論

テクニカル

- DAC/トランシーバ/NICの品質問題とその対処
- DCNW内でのIP Anycastにおけるトラブル、トラブルシューティング
- IPv6の利用の推進、無理にIPv6使わなくてもいいじゃないかというご意見への対処
- ネットワークコントローラ、分散 vs 集中（どういう機能分担がよいか）
- SONiC/ホワイトボックスの活用方法、ホワイトボックス時代に求められるスキルの変化（NWエンジニアもLinuxの知識が必要になる）

非テクニカル

- 「巨大プロジェクト」との向き合い方（ステークホルダ、レポーティング）
- 挑戦や変化に(伴い問題が発生することに)対するネガティブな反応
- ブラックボックスな製品の中で起こる問題への向き合い方
- 人材育成、内製への変化（特にネットワークエンジニア）

重点議論ポイント①

・理想はテナント要件確定後、物理側の要件定義＆調達を実施する事
・一方で「HW調達・構築」が最も時間がかかる工程でもある
　⇒テナント要件確定する前にある程度、物理側を作り始めることも致し方なしに思える

いい解決策などあれば教えてください

マルチサイトソリューションの破綻④

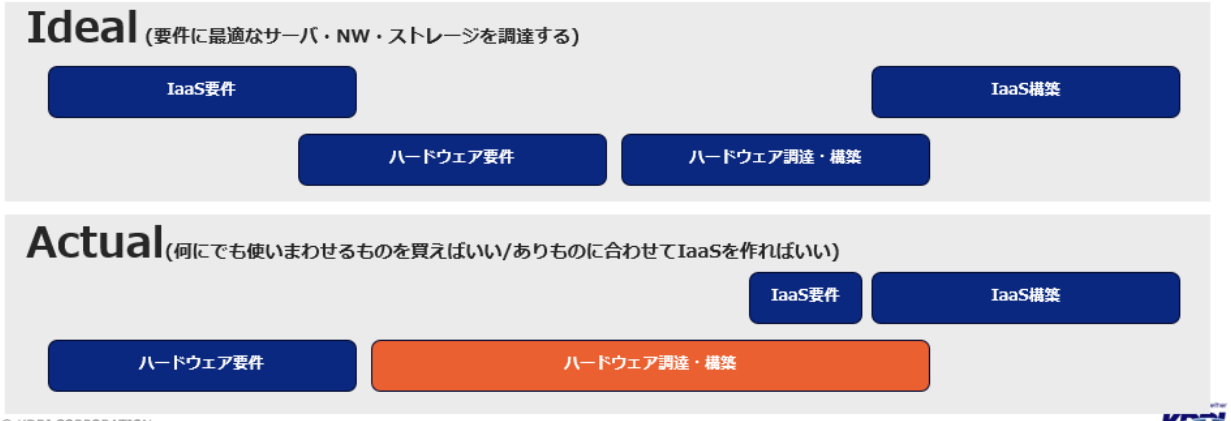
メリット
- L2でregion間を延伸可能
- プロビジョニングが迅速
  - コントローラからワンタッチでSite間通信を延伸可能

導入後、上記をテナントが求めていないことが判明

重点議論のポイント①

＋あいまいなまま進んでいたこと：IaaSの設計に先行してHW調達

リードタイムの長期化等によりHWを先行調達
IaaS/プライベートクラウドは後追いで作ることに



重点議論ポイント②

DACケーブル

- DACケーブルでの 400G -> 100G x4 Break-out の利用実績ありますか？
- Input error(CRC error)をどの程度気にしていますか？
  - KDDIの環境では結構クリティカルなUDPパケットがたくさんあるので気にしてしまう
- ベンダ様からはBreak-outでの混信・電磁干渉の可能性の示唆がある状況だが決定的な証拠等がある状況じゃない

エンジニア育成

- 内製化が基本のIaaSチーム
- パートナー様への発注が基本のNWチーム

- Linuxでオープンソースソフトウェアを見るIaaSチーム
- ブラックボックス機器を相手にしなければならないNWチーム

重点議論ポイント③

アーキテクチャ見直し

- スモールスタート・PoCが基本だとは思いますが…一度入れて微妙だった時の見直し方
- 一度決めたことを覆すことへの心理的抵抗や心理的負荷

議論

テクニカル

- DAC/トランシーバ/NICの品質問題とその対処
- DCNW内でのIP Anycastにおけるトラブル、トラブルシューティング
- IPv6の利用の推進、無理にIPv6使わなくてもいいじゃないかというご意見への対処
- ネットワークコントローラ、分散 vs 集中（どういう機能分担がよいか）
- SONiC/ホワイトボックスの活用方法、ホワイトボックス時代に求められるスキルの変化（NWエンジニアもLinuxの知識が必要になる）

非テクニカル

- 「巨大プロジェクト」との向き合い方（ステークホルダ、レポーティング）
- 挑戦や変化に(伴い問題が発生することに)対するネガティブな反応
- ブラックボックスな製品の中で起こる問題への向き合い方
- 人材育成、内製への変化（特にネットワークエンジニア）

Tomorrow, Together

KDDI

「つなぐチカラ」を進化させ、
誰もが思いを実現できる社会をつくる。

KDDI VISION 2030